0

HARVARD UNIVERSITY

DEPARTMENT OF STATISTICS

TEL. 617-495-5496
FAX. 617-496-8057

SCIENCE CENTER
ONE OXFORD STREET
CAMBRIDGE, MASSACHUSETTS 02138

Issues in DNA Fingerprinting

Herman Chernoff

Harvard University

Technical Report No. ONR-C-16

July 1994

24pg

Abstract

The use, in court, of DNA Profiling, popularly referred to as DNA Fingerprinting, for forensic identification purposes has been questioned. A report of the National Research Council was solicited to clarify the issues and propose procedures of how and where this powerful technique could be used. This report has been subject to criticism. The main point of the report was to recommend procedures for the primary issue of quality control and for gathering useful data. The report also proposed the *ceiling principle* as a conservative approach for calculating the match probability for an innocent suspect to be used in assessing the guilt of a suspect. That method has been criticized by some as not necessarily conservative, and by others as unnecessarily conservative as well as illogical. These issues are discussed as well as some of the recent history in court.

AMS 1991 Mathematics Subject Classifications: Primary 62P25; Secondary 62A16, 62P10.

Key words and phrases: DNA fingerprints, DNA profiles, Bayes theorem, loci, Frye Rule.

Issues in DNA Fingerprinting*

Herman Chernoff
Harvard University

1. Introduction

The use of blood types for identification of parents and criminals began to be replaced in 1985 by the much more powerful technology of *DNA Profiling*, popularly referred to as *DNA Fingerprinting*, (Jeffreys *et al.* 1985). Since then the use of this method in court has been questioned, and it has been the subject of numerous discussions in legal and scientific journals.

We have a technical tool, with a power orders of magnitude greater than what was available earlier. It is incomparably better than eyewitness identification by strangers which is often admitted as evidence. Still the resulting evidence has on occasion been discarded by judges because expert witnesses for the defense denied that the claimed odds of many billions to one were reliable (Kaye 1993, p. 118).

It is likely that in the near future most courts will admit such evidence, and the technology will be refined, but the debate about applicability will continue for some time. One important saving grace of the current situation, is that DNA profiling is extremely effective in screening out innocent suspects. Where well preserved biological specimens belonging to the guilty party are available, it is almost impossible for an innocent person to be falsely identified with an honest and careful analysis. Thus, when the technology is applicable, innocent suspects are likely to be recognized quickly, saving them considerable pain and eliminating false leads in police searches. The method has been used successfully to establish the innocence of prisoners who were convicted of rape years before DNA profiles became widely available, (Kaye 1993, p.115).

There are several issues which are not easy to resolve. Some of these stem from the adversarial nature of the legal system where it is the function of a lawyer to present the best possible case for his client. It is not the lawyer's function to give a balanced view of the evidence. Rather, the defense must seek to create doubts, sound or otherwise, among the judges about the admissibility, and jurors about the weight, of the evidence presented

* Research supported in part by ONR Contract N00014-91-J1005 and NSF Contract DMS-9104990

by the prosecution. In an area where arguments are seldom quantified, probabilistic reasoning is not well understood, and the underlying models used by statistical analysts are approximations, odds of billions to one are easy to criticize. The result is often unnecessary confusion, which a good defense lawyer can exploit.

To put it bluntly, a false identification is much more likely to be due to an error in labeling the biological specimens, or tampering with the evidence, or even to the existence of a hitherto unknown evil identical twin, than to a "false match" of the specimens.

Although the courts and technology will probably arrive at a moderately sensible and stable equilibrium in the treatment of DNA profiles within the next few years, the history is a reflection of the uneasy tension derived from the use of expert witnesses in legal matters involving the combined professions of law, some branch of science or technology, and statistics. It may be that a discussion of the history and the issues will contribute enough clarity to help reduce somewhat the trauma, in putting into practical legal use the next technological breakthrough, for establishing relevant matters of fact.

2. DNA Analysis

The power of DNA analysis derives from the stability of the DNA molecule, and the uniqueness, except in the case of identical twins, of an individual's DNA. With the exception of sperm cells in men and egg cells in women, the nucleus of each living cell in the human body carries 23 distinct pairs of *chromosomes*. The sperm and egg cells each carry only one set of 23 chromosomes. Of each pair of chromosomes, one comes from the father's sperm and one from the mother's egg. Each of the 46 chromosomes is a molecule consisting of a long strand of DNA in the form of two intertwined substrands.

The chromosomes are duplicated in the process of cell reproduction, and parts of each pair are combined to form a single chromosome in egg and sperm production. Because the chromosomes carry the essential instructions for cell function, this duplication process is almost error free, and changes, called *mutations*, rarely survive. Some surviving mutations lead to apparent changes in appearance, but others, at certain sites or *loci* of the chromosomes, seem to have no effect. Certain loci are much more predisposed than others to tolerate mutations, and these are highly *polymorphic* giving rise, in the population, to many distinct forms called *alleles*, which are detectable in the laboratory. Inasmuch as they seem to have no apparent function, these *silent* polymorphisms do not affect mating behavior.

One laboratory technique cuts out a section of each of a chromosome pair at a certain locus and effectively weighs the sections to distinguish the various alleles. The result is a column with two crude rectangular shadows or bands. The locations of the bands in the column are subject to measurement error and are called *band weights*. Sometimes only one band appears. This may be because the subject is *homozygous* (identical alleles) or two bands *coalesce* because the weights are so similar, or one band weight is off scale, or the specimen has degraded. The results for the forensic specimen and the suspect are compared in adjacent columns, visually or by machine. A match is declared if the bands for the two specimens line up closely enough.

If there are 30 alleles for a locus such as D12S79, and each individual has two chromosomes for this locus, there are $465 = 30(31)/2$, not necessarily uniformly distributed, possible variations. The number of possibilities becomes enormous if several different loci are used.

There are other established methods of discriminating between alleles. Each is associated with advantages and disadvantages. These involve costs, errors in measurement, difficulties with forensic specimens which may be degraded and in small quantities, and contamination with extraneous DNA. One such promising procedure is the use of PCR (polymerase chain reaction) which multiplies enormously the amount of DNA material from a portion of a specimen. It is said to have the ability to provide as much DNA as necessary from a single nucleated cell. It is quick and economical. It is also easily contaminated, and forensic experience with it is limited at present. There are available several techniques for determining alleles with PCR.

3. Statistical Considerations

Statistical considerations enter primarily in two ways. Because the match decisions for band weights are subject to measurement error, some allowance must be made. Even assuming no errors in measurement, the presence of a match must be evaluated for its weight of evidence in favor of the hypothesis that the two specimens are from the same individual.

Let us ignore the error in measurement problem until Section 8. A typical calculation follows. Supposing that the forensic specimen comes from a guilty party, what is the chance that an innocent suspect will provide a match with the two measurements ordinarily derived from the forensic specimen at a given locus. If the distribution of weights in the population assigns probabilities p_1 and p_2 to the observed forensic band weights (which we will assume here are different) then the probability that a randomly selected innocent subject will lead to a match is calculated, using statistical independence assumptions, to be $2p_1p_2$. If p_1 and p_2 are about 0.05 each, this product is 0.005. Moreover, if we obtain matching in four independent probes with comparable discriminating power, we will have observed an event with probability of the order of magnitude of one in a billion.

As we shall see, this calculation is subject to criticism. Moreover, the probability calculated here does *not* purport to be the probability that the suspect or an identical twin was responsible for the forensic specimen. The evidence is most easily assessed in terms of Bayes Theorem. Let G and I represent the alternative hypotheses that the suspect is and is not responsible for the forensic specimen. Allowing for the possibility that the forensic specimen might have been left by an innocent party, the notation of G and I for guilt and innocence is slightly misleading, but mnemonically helpful. Let E be the evidence observed and we may have $E = M$ for *match* and $E = D$ for different or *nonmatching.*

The approach using Bayes Theorem requires some *prior probabilities* $P(G)$ and $P(I) = 1 - P(G)$ which will depend on the juror's evaluation of other information. Then $P(G)/P(I)$ is called the juror's *prior odds* in favor of G. We introduce the *posterior probabilities* $P(G|E)$ and $P(I|E)$ and the *posterior odds* $P(G|E)/P(I|E)$ which represent the juror's probabilities and odds *given the evidence* E. Finally the *likelihood ratio* is $P(E|G)/P(E|I)$, the ratio of the probabilities of the evidence given the hypotheses.

Bayes Theorem states that the *posterior odds is equal to the prior odds multiplied by the likelihood ratio.* Thus

$$\frac{P(G|E)}{P(I|E)} = \frac{P(G)}{P(I)} \cdot \frac{P(E|G)}{P(E|I)} .$$

In other words the likelihood ratio is a factor by which the juror's prior odds is converted to his or her posterior odds. In our problem we assume that if $E = D$ (nonmatch), then $P(D|G)$ is small while $P(D|I)$ will be quite close to one. In that case, a nonmatch will multiply the prior odds by a very small number and tend to overwhelm, in favor of I, any prior evidence. On the other hand if $E = M$; then $P(M|G)$ is close to one while $P(M|I)$ is the small number we indicated previously to be about one in a billion or 10^{-9}. Then the likelihood-ratio is huge and the prior odds will be multiplied by a large number supporting a large increase of the posterior odds over the prior odds. For example if the prior odds were $1/4$, or 1 for G to 4 for I, a match with $P(M|I) = 10^{-9}$ would give posterior odds of about 250 million to one, and the posterior probability of I, $P(I|M)$ would be approximately $4 \cdot 10^{-9}$ and not the same as $P(M|I)$ or 10^{-9}.

For the benefit of readers not accustomed to statistical reasoning and Bayes theorem, an example which might lend some additional perspective is a simplified version of that of HIV testing with an instrument which never gives a false negative, but gives false positives 5% of the time (for disease free subjects) on a population with 1% infected. Many individuals who tested positive thought that they were infected and some felt that their probability of being infected was 0.95. By Bayes theorem the posterior odds of infection given the evidence was the prior odds times the likelihood ratio or $(0.01/0.99) \times 1.0/(0.05) \approx 0.202$. This means that the posterior probability of infection is close to 0.20 and not 0.95. In other words, while $P(+|\text{diseasefree}) = 0.05$, $P(\text{diseasefree}|+) \approx 0.80$. Note that the assumed *prevalence rate* in the population, 0.01 , analogous to the prior probability of G in our forensic example, is extremely important here.

In our forensic example above the distinction between $P(I|M) = 4 \times 10^{-9}$ and $P(M|I) = 10^{-9}$ seems unimportant since either number is overwhelming. Any scientific problem in drawing conclusions must come from questions about how the probability 10^{-9} was calculated in the first place. However there also exists the practical problem of how to present such evidence to a judge or jury so that it is meaningful.

In this context, attention has been drawn by Thompson and Schumann (1987) to the "prosecutor's fallacy" and the "defense attorney's fallacy." The court is primariliy interested in $P(I|M)$ or $P(G|M) = 1 - P(I|M)$. The "prosecutor's fallacy" is to confuse $P(M|I)$ with $P(I|M)$. The HIV example shows that the distinction may sometimes be important. The "defense attorney's fallacy" is to argue if $P(M|I) = 10^{-5}$, that we live in a city of 10^6 people and thus there are about 10 people who match and so $P(G|M) = 1/10$. Implicitly the defense attorney is suggesting that prior to the evidence of matching, everyone in the city is equally likely to be the source of the forensic specimen and the prior odds should be $1/10^6 = 10^{-6}$. If the suspect were there because there were other signs of guilt, the juror has a right to believe, on the basis of these other signs, that the prior odds ought to be substantially greater than one in a million.

4. Legal Considerations

One of the first uses of DNA profiling, described in Wambaugh (1989) came up in a village in England where two 15 year old girls were raped and murdered three years apart. The DNA analysis of semen found on these two bodies determined that one perpetrator had been involved in both cases. The plan, to test every man of ages between 17 and 34 within a few miles of the village, failed to reveal a match because the perpetrator convinced an acquaintance to substitute his own blood sample for the perpetrator's. When that evasion was uncovered, the perpetrator confessed.

Although a subsequent test revealed a match, it is ironic that that evidence was unnecessary, because the perpetrator, convinced of the accuracy of the test and admissibility of the evidence, confessed. Because several thousand men were tested, the appropriate version of the defense attorney's fallacy would have been relevant, reducing the effect of the evidence on the posterior odds of guilt in this case.

The immediate reaction to the introduction of DNA profiles in legal circles was one of enthusiasm for a major breakthrough. This was tempered by warnings, described by Thompson and Ford (1989), against the premature introduction in court of this approach, before the tests have been adequately validated and debated in the scientific literature, leading to general acceptance in the scientific community. Attention was drawn to another method, protein gel electrophoresis, similar in some aspects to DNA profiling, but lacking its highly polymorphic property. Its use had become routine, but was subsequently ruled inadmissible in California and Michigan after serious doubts were raised about its reliability. This development raised havoc in law courts because the method had provided important evidence in a large number of cases.

Certain safeguards have been generally accepted by the courts to deal with expert testimony in general, and testimony involving novel technological develpments in particular. In response certain safeguards have been generally accepted. The most frequently cited rule for determining the *admissibility* of a novel scientific technique is the Frye Rule (Kaye 1993, p.104). Under this rule the judge must decide that the "thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs." Courts have noted that the particular procedure need not be "unanimously endorsed" by the scientific community, but must be "generally accepted as reliable." The use is questionable if there is no pool of experts which would permit the opposing side to find experts who can critically examine the validity of an opponent's scientific evidence. Under the Frye rule it is not necessary to reopen the issue of admissibility in each case, but considerations of *weight of evidence* may have to be considered by individual courts.

The frustration with the delay between the development and acceptance of new technology has led some jurisdictions to apply a more liberal approach for admissibility. Rule 702 of the Federal and Uniform Rules of Evidence requires only relevance, in that qualified scientific testimony should assist in understanding the evidence or to determine a fact in issue. On the other hand, Rule 403, which applies to all evidence, scientific and non-scientific alike, provides that "evidence may be excluded if probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury." This rule has concerned courts faced with probabilities of 1 out of several hundred billion for fear that jurors may be overly impressed by such numbers and give them undue weight and deference, (Kaye, 1993, p.118). This does not mention some, possibly well placed, cynical skepticism about the meaning of such numbers.

A recent decision by the Supreme Court (Kaye 1993, p.105) has established the precedence of the Federal Rules of Evidence over the Frye Rule for Federal cases. Here, as in many judicial opinions, the term *scientific validity* tends to appear in a somewhat circular way.

5. Problems

The warnings against premature use of DNA profiles were well founded. By 1989, enough questions were raised that the scientific and legal communities solicited a report by the National Research Counsel (NRC) of the National Academy of Sciences. Unfortunately this report, (Committee on DNA Technology in Forensic Inference 1992), has received considerable criticism and has not quite served as well as intended to settle the issues.

Three commercial firms, Lifecodes, Cellmark and Cetus, as well as the FBI and many state agencies have been involved in providing forensic analysis. One of the difficulties at the beginning was the lack of a pool of qualified experts for use by the defense. That was partly due to the fact that expertise was required in three distinct areas. Laboratory scientists could testify about the technology. Academic geneticists and statisticians could testify about probability. Neither of these two groups have much contact with specialists in forensic science, where one must learn to deal with degraded specimens that are not obtained from well run scientific laboratories. Some of the expert witnesses were not very expert in any of these fields. Also troublesome was the fact that many of the experts for the prosecution represented the commercial laboratories doing the analysis and had some incentive to build up their business by claiming how effective their techniques were. Moreover there was little significant statistical expertize in these laboratories.

A particularly difficult point was that the commercial laboratories were reluctant to share their techniques and proprietary data, mostly derived from opportunistic rather than random samples, with potential competitors. How could a defense expert criticize a calculation based on estimates of certain probabilities derived from unspecified models and unavailable studies of the distribution of polymorphisms in the population? Geisser (1990) tells of a defense lawyer who cleverly used this reluctance to provide the data, to have the evidence ruled inadmissible.

In addition Geisser states that the FBI was also reluctant to cooperate with scientists requesting data. which led to the embarrassment of the FBI of having a scientific publication point out the difficulty of accessing the data on the basis of which their estimates were calculated. Finally some of the scientists acting as expert witnesses for the defense claimed harassment on the part of authorities ranging from interference in scientific publications to questions about visa status (Kolata 1991).

In 1989, the Castro case in New York, attracted a good deal of attention. Highly qualified experts for both sides were brought in from all parts of the country and they produced 2,000 pages of testimony. In this case, an error in laboratory work left the issue of probability estimates moot. This case was subsequently the illustrative example of Berry (1991) where it was pointed out that Lifecodes had apparently underestimated the measurement error in their procedures. The measurement error determines a standard deviation for the difference of the measured band weights of an allele for two specimens from the same individual. Two band weights are said to match if their difference is less than k estimated standard deviations of the difference. Over time Lifecodes changed k from 2 to 3. But in the Castro case, even though one of the differences was 3.66 standard deviations, Lifecodes decided to call it a match. By underestimating the measurement error, and consequently the standard deviation of the difference, the people at Lifecodes put themselves in the embarrassing position of stretching the criterion in order to claim a match, thereby destroying a pretense at objectivity.

It was hoped that the NRC report would establish a clear cut explanation of how, when and where DNA profile analysis could be used, and that this report would be accepted by the scientific community. Instead it gave rise to considerable criticism (Aldhous 1993). In response to this criticism, the NRC is planning to produce a second report to clarify ambiguous wordings and issues on which they may have left chinks in the armor against wily defense lawyers and prosecutors.

6. Recommendations of the NRC Report

The main issues addressed by the NRC report were quality control in a rapidly evolving technology, objectivity and scientific standards. To deal with another major issue, the weight of evidence, the *ceiling principle* was proposed as a conservative method for dealing with the calculation of the probability of a match $P(M|I)$.

Some recommendations involve precautions for dealing with new developments and future standards. These include publication, accreditation, and the use of data to be gathered by sampling the population for blood specimens, which should be stored for further analysis as new techniques develop. For current methods, recognized as fundamentally sound, open availability of data and laboratory records were stressed.

One problematic recommendation was "Prosecutors and defense counsel should not oversell DNA evidence. Presentations that suggest to a judge or jury that DNA typing is infallible are rarely justified and should be avoided." Another follows the *ceiling principle,* to be described later. "Until ceiling frequencies can be estimated from appropriate population studies"..."The testing laboratory should then calculate an estimated population frequency"..."provided that population studies have been carried out in at least three major 'races' (e.g., Caucasians, blacks, Hispanics, Asians, and Native Americans) and that statistical evaluation of HW equilibrium and linkage disequilibrium has been carried out"... The report does not say what to do if the tests reject equilibrium. The concepts of HW and linkage equilibrium will be discussed in Section 8.

7. Criticism of the NRC Report

A newspaper article published just before the NRC report was made public claimed that the report recommended that DNA evidence be barred from the courts (Kaye 1993, p.103). In response, Victor McKusick, chairman of the NRC committee, held a press conference stating that the report "approved of the forensic use of DNA substantially as it is now practiced." In a book review Lewontin (1992) quotes several portions of the report which seem to be less than wholehearted in support of current practice.

It is his contention that it is in the nature of NRC reports to be unanimous, and so reports should be expected "to contain contradictory compromises among contending interests." The portion of the recommendations that warns against the infallibility of DNA typing seems to be an example of such a compromise. While everyone may agree that low quality analysis might lead to misleading conclusions, this sentence, vague as it is, could conceivably be used as a wedge to break open an otherwise tight prosecutor's case.

Lewontin quotes several other portions of the text of the NRC report. One example is "The current laboratory procedure for detecting DNA variation... is *fundamentally* sound. [emphasis added]." Thus this sentence can be regarded as less than a ringing endorsement. Another example consists of three portions "It is now clear that DNA typing methods are a most powerful adjunct to forensic science for personal identification and have immense benefit to the public" and later "DNA typing is capable, *in principle*, of an extremely low inherent rate of false results [emphasis added]" and these are followed by "The committee recognizes that standardization of practices in forensic laboratories in general is more problematic than in other laboratory settings; stated succinctly, forensic scientists have little or no control over the nature, condition, form, or amount of sample with which they must work." It isn't clear that these passages are as self defeating or damning as Lewontin suggests, but they do present difficulties for judges in individual courts.

In my opinion the really important issue is that of quality control which was well addressed by the report. There is no way that the report could, without sounding tentative, point out that we have an immensely powerful tool that needs a great deal of care and expertise to be applied properly. However the issue which is the source of the most criticism in the scientific literature is the ceiling principle for the calculation of $P(M|I)$. Here the criticism comes from both the right and the left if I may use these politically loaded terms to represent conservative and not conservative, in the sense of favoring or not favoring the suspect. On one hand Cohen (1992) demonstrates that there are possible circumstances where the principle is not conservative. On the other hand, it is attacked as being unnecessarily conservative as well as illogical.

What is the potential harm of being too conservative? Then the prior odds is multiplied by a much smaller number than is appropriate to calculate the posterior odds of

G to I. Starting with a prior of 1/10 a likelihood ratio of 10^5 in place of 10^8 would result in a posterior of 10,000 to 1 in place of 10 million to one. There is no problem in going from the evidence M to G, the hypothesis that the specimens, allegedly from the crime scene and from the suspect, are from the same source. When one considers the apparently common practice of tampering with the evidence, or the possibility of an accidental mislabeling of the specimens, or in the case of PCR, the contamination of the specimens, the weak link, if any, is in going from that hypothesis G to actual guilt.

To be sure there may be exceptional cases where the likelihood ratio is not so compelling. Typically such cases are those which don't follow the usual routine and demand real expertise for proper analysis. The NRC emphasis on quality control and training and certification for expert witnesses is needed to deal with those cases which seem to appear with surprising frequency, considering the adjective exceptional.

8. The Ceiling Principle

A fundamental result in genetics (Weir 1990) involves a large population in which there are k possible alleles at a certain locus in proportions $p_1, p_2, \ldots, p_k$. After one generation of *random mating* within the population, these proportions will not change much, and the proportion of the population which have the i-th and j-th alleles will be approximately $2p_ip_j$ for $i \neq j$ and p_i^2 for $i = j$. This corresponds to the statement that for a randomly selected offspring, each of the two alleles he or she inherits are independently selected with probabilities p_i for $i = 1, 2, \ldots, k$. Such a population is said to be in *Hardy-Weinberg (HW) equilibrium.* The assumption of random mating is most likely to be accepted for those loci for which the polymorphisms are silent.

If we deal with a stable population which has few people coming in or out for several generations, and a locus whose alleles have no outwardly observable distinct manifestations, one might expect HW equilibrium to hold. Most populations that are sampled to estimate the p_i are not of the type described above. For example U.S. Hispanics represent mixtures of quite separate subgroups, some of which may be thought of as relatively homogeneous. We show what can happen with a simple artificial example. If three alleles $1, 2, 3$ appear with frequencies (0.3, 0.5, 0.2) in one subgroup and with frequencies (0.5, 0.1, 0.4) in a second subgroup of equal size, the combined group will have frequencies (0.4, 0.3, 0.3). Assuming HW equilibrium within each subgroup and no mating across these two subgroups, the frequency of offspring with alleles (1,2) would be $(0.3)(0.5) + (0.5)(0.1) = 0.20$. Treating the combined group as though it were in HW equilibrium, we would calculate the wrong result $2(0.4)(0.3) = 0.24$.

In the formation of the egg and the sperm, a pair of the parent's chromosomes may undergo a process where each chromosome may break into several relatively large pieces which recombine with complementary pieces of the other. The *recombination* takes place so that the egg or sperm has one complete set of 23 chromosomes, but each of these may have sections from each of the originating pair. Thus, for example, each chromosome of the sperm may carry part of the chromosomes originating from each of the father's parents. Alleles of two loci that are close to each other on a chromosome are likely to be *linked* in the sense that if one appears in the egg or sperm, the other is very likely to accompany it. For loci that are far apart on a given chromosome, there tends to be practically no such linkage.

For loci on different chromosomes there is believed to be no such linkage. It follows that for a stable population with random mating the alleles of loci on different chromosomes are statistically independent. This is called *linkage equilibrium* and permits one to multiply probabilities of events involving loci on different chromosomes to obtain resulting probabilities for the occurrence of all the events. Neither HW equilibrium nor linkage equilibrium is expected to obtain for populations consisting of several different subgroups, among which there is not much mixing, i.e., where random mating does not apply. The existence of such subgroups is referred to as *population substructure.*

If HW and linkage equilibrium are assumed, the probability of a random individual matching a given profile may be calculated as the product of terms $2p_{i1}p_{i2}$ where p_{i1} and p_{i2} are the relative frequencies of observed alleles i_1 and i_2 of the ith locus in the *heterozygous* case where these alleles are different, or of terms p_i^2 where p_i is the probability of the observed allele i in the *homozygous* case where both chromosomes have the samc allele i. The practice of assuming HW and linkage equilibrium when there is population substructure will lead to incorrect estimation of $P(M|I)$. However experiments based on artificially combining various groups have consistently shown relatively minor effects due to assuming independence when substructure exists. For this reason it is conjectured that a test for lack of HW equilibrium will have very little *power* (ability to reject the hypothesis of equilibrium) without very large sample sizes. Lack of power is correlated with a relatively minor effect on the calculation of $P(M|I)$.

Considering the lack of power of the test, it should be remarked that various investigations had no difficulty in detecting lack of equilibrium by observing more homozygous cases than expected. Others, e.g. Devlin *et al.* (1990) have attributed that effect to coalescence and measurement error.

When only one band appears for the specimen, there are other possible explanations besides homozygosity and coalescence. One is that one weight was so small or large that it was offscale for the measuring instrument and did not appear. The other involves degraded specimens. To deal with the first explanation the NRC report recommends replacing p_i^2 by $2p_i$ which is considerably more conservative.

Several treatments of the measurement error problem are referred to by the terms *floating bins* and *fixed bins.* We will describe the fixed bin approach used by the FBI. The possible observations on band weight are divided up into cells or nonoverlapping bins. Several populations including Caucasians, Hispanics and Blacks are sampled separately. For each population sampled and each locus used, a histogram is formed to estimate the frequency distribution of band weights, i.e., the proportion of the measurements that fall into each of the bins.

The band weights of a forensic specimen and a suspect are declared to match if the difference in the observations is small enough, i.e., less than 5% the average weight. Weir and Gaut (1993) state that in studies conducted by the South Carolina Law Enforcement Division (SLED) DNA laboratory, corresponding bands, from a blood sample and a vaginal swab (containing epithelial cells) from each of a sample of a women, never differ by more than 5.6% of their average band weight. Consequently, at SLED a wider matching interval is used than at the FBI, and a nonmatch is declared when the difference in corresponding bands exceeds 5.6% of the average length. In both labora'ories, when a nonmatch is declared, the suspect is cleared. If the suspect is not cleared on any of the loci tested, a match is declared and a probability $P(M|I)$ is calculated and associated with the event of a match.

According to the ceiling principle, the estimate of $P(M|I)$ is then a product of terms of the form $2p_{i1}p_{i2}$ for heterozygous cases at the ith locus and of the form $2p_i$ for apparent homozygous cases. The value of the p corresponding to a band weight for a locus depends on the position of the interval $e(1 \pm 0.05)$ where e is the band weight of the forensic specimen. If this interval lies within one of the fixed bins, use a conservative estimate of the probability of falling in the fixed bin as p. If the interval overlaps two or more of the fixed bins, the NRC report recommends adding the estimated probabilities of the overlapped fixed bins (I assume here that they mean to have a conservative estimate of the probability of falling in any one of the overlapped fixed bins.)

This proposal was criticized by Weir and Gaut (1993) among others who argue that one should only consider the maximum of the probabilities for the overlapped fixed bins. Since the fixed bins are usually considerably wider than $2(0.05)e$, then either one or two bins are overlapped. A fixed bin may contain several alleles, each of which contributes some part of the probability of falling in that bin. If most of the probability corresponding to each of two neighboring bins is concentrated on alleles close to the common boundary, then the NRC proposal is essential for an interval overlapping these bins. Weir (personal communication) argues that a study of the data in the FBI samples indicates that this contingency does not arise.

The ceiling principle involves estimating each bin probability by an appropriate upper bound. This is to be selected by sampling various data bases. "Random samples of 100 persons should be drawn from each of 15-20 populations, each representing a group relatively homogeneous genetically; the largest frequency in any of these populations or 5% whichever is larger, should be taken as the ceiling frequency."..."The goal is not to ensure that the ethnic background of every particular defendant is represented, but rather to define the range of allele frequency variation." The report also adds an interim suggestion while the samples are being collected of (1) using 10% in place of 5%, (2) using upper 95% confidence bounds on the bin frequencies for the separate racial groups studied to date, and (3) indicating how rare matches are by mentioning the total number of profiles in the combined data bases on the assumption that none of them match the forensic specimen.

9. Probability Model

Ideally we wish to see which of the possible perpetrators lead to a match with the forensic specimen. In general it is not possible to identify and assemble such a group. Even in the original case cited (Wambaugh 1989) the group tested did not include potential perpetrators who were older than 34 or from outside the local area. However one may ask how rare is this matching profile.

The natural method for making a quantitative evaluation is to construct a *relevant* probability model from which a reasonable, or at least a conservative estimate of $P(M|I)$ could be calculated. Our discussion suggests a model where the suspect is selected at random from a *reference population* of possible perpetrators, and $P(M|I)$ would be the proportion of that population which has matching profiles. But, as indicated above, not enough is ordinarily known about the profiles of that population to be useful.

One conservative approach is to find a substantial population, in which the forensic profile or the alleles of that profile are very common. Then we might ask what is the probability that our suspect is a randomly selected innocent individual, from that population, who happens to have a profile matching that of the forensic specimen. Another conservative approach is to propose the population from which the suspect comes and to estimate $P(M|I)$ as the probability of finding a matching profile from that population.

The last method is conservative in the following sense. Without looking at the suspect's profile, the expectation of the log of the probability of a match with someone of an arbitrary population is no greater than that with someone of the suspect's population. However the suspect is a member of several populations. For example we could take all his close relatives including his siblings if he has any. That population may be difficult to use because the relatives may be unwilling to be profiled, and besides there may be no such relatives who are possible suspects.

It may be that the suspect is a member of a relatively homogeneous group on which there exists data and for which geneticists may feel satisfied that the loci considered are such that there is every reason to expect genetic equilibrium among the loci studied. That would constitute a reasonable group to use to provide a conservative estimate of $P(M|I)$ using independence.

The NRC report attempts to bypass the reference population issue by sampling a substantial number of relatively homogeneous populations to see how much allele frequencies vary among these, and to replace estimated frequencies by upper bounds independent of the particular population.

To return to the population of relatives, one might construct a hypothetical infinite population of possible siblings of the suspect. That would provide a very conservative estimate of $P(M|I)$ based on HW and linkage equilibrium, and every locus where the suspect is heterozygous would provide a factor a little smaller than 4 (Weir and Hill 1993). Then 4 loci could contribute a likelihood ratio of no greater than $4^4 = 256$. To achieve more impressive likelihood ratios it will be necessary to use more loci. Such a scheme was proposed by Belin (personal communication). One advantage of such a scheme is that HW and linkage equilibrium is more likely to be accepted in this context by geneticists. Another is that estimates of matching probabilities at individual loci, from empirical data cumulated from siblings, would not be sensitive to the populations from which the samples were drawn or even how the samples were drawn.

The last proposal involves a hypothetical reference population. To apply HW and linkage equilibrium in other populations, we are more or less forced to invoke a hypothetical infinite population from which the individuals in our sample and the suspect are drawn. The reason is that once we have matching on 4 or 5 loci, in a population of a few billion, our individual has been almost certainly uniquely identified, in which case we are almost sure to have matching on the other loci. In other words, for finite populations, independence breaks down for events involving matching at several loci. However this breakdown is not a basis for critics of DNA profiling to cite as evidence favoring the defense. For homogeneous subpopulations most geneticists seem willing to accept HW and linkage equilibrium as a reasonable model, at least for loci in different chromosomes.

What I have attempted in this section, is to point out some fundamental problems in defining a suitable quantitative measure of $P(M|I)$ and to suggest that the NRC approach is very conservative even though it invokes HW and linkage equilibrium.

10. Criticisms of Ceiling Principle

Cohen (1992) has demonstrated, with an artificial counter-example, that the NRC claim that their approach is conservative, even in the presence of substructure, does not hold up and can fail badly. His example depends on lack of linkage equilibrium within the subpopulations. However, experiments carried out, which created substructure in samples by combining data from different real populations, tend to reveal little effect from assuming independence (Devlin *et al.* 1993).

The NRC report has received criticism from those who feel that the ceiling principle is too conservative. It is argued that science demands a good estimate and not an overly conservative estimate. But for the decision makers, the issues are guilt or innocence, and the point of the DNA evidence is to establish whether or not it is reasonable to believe that the suspect was almost surely the source of the forensic specimen. Whether a $P(I|E) = 10^{-4}$ or 10^{-8} is usually irrelevant, because judges, when sampled, regard probabilities from 0.75 to 0.95 as "beyond a reasonable doubt" (Gstwirth,1988).

Another argument (Weir and Gaut 1993) is that the ceiling principle suggests using probabilities of exclusive events which don't add up to one. This violates the laws of probability. As an undergraduate, my school mates and I played a game, the object of which was logically to derive a given statement from $1 = 2$ as quickly as possible. That this is possible is the reason that applied mathematicians must check their models for coherence, existence and uniqueness before engaging in complex analysis. But the calculation using the above violation of probability is simple and transparent and is clearly dedicated to provide a conservative estimate of $P(M|I)$ in applications where the effect of substructure are expected to be minor. The danger of serious errors, other than conservatism, from the application of the ceiling principle, is negligible under treatment by experts.

There are arguments (Devlin *et al.* 1993) based on the data collected and experiments with those data that indicate that the ceiling principle is unnecessarily conservative. These arguments are persuasive, but they are based on a few experiments and limited data. It is claimed that overly conservative estimates require the use of many loci, and this increases the probability of the other error, i.e., that of incorrectly declaring a mismatch. But, that deficiency can easily be countered by increasing slightly the width of the interval of bandweights used to declare a match. As a result the use of additional loci might be a little less effective than expected by a naive analysis based on the original intervals. However, using more loci provides more information and can be used to reduce both the probability of declaring a match inappropriately and of declaring a mismatch inappropriately.

(Devlin *et al.* 1993) claim that the plan to sample 100 individuals from each of a dozen populations is not a good experimental design and will force unduly conservative estimates. Finally (Kaye 1993, p. 172) the choices of 5% and 10% "rest on an unarticulated balancing of competing policies." But with four loci the 10% figure is still adequate to obtain enormous likelihood ratios. With fewer than 5% of the population at a locus, the estimate of that proportion in samples of 100 could fluctuate wildly. No sharp analysis is required to see that these choices are adequate, if not optimal, for a conservative, yet effective approach.

11. Philosophy of Science

A major shortcoming, in the argument that the effect of assuming independence is minor, is that the evidence is based on a few experiments with a limited number of data sets with few subjects, and often with opportunistic samples consisting of subjects who were readily available. One is inclined to believe that these results will generalize widely, but such an inclination is not proof.

Scientists typically do not deal with problems demanding near certainty even though lawyers may feel that the scientific method gives clear and definitive results. The reality is somewhat different. Scientists perform experiments suggested by partly baked hypotheses, conjectures and models to gather data, the analysis of which lead to better baked hypotheses and sharper experiments. At the end of a series of such experiments, a murky mess of ideas will gradually have given way to a clearer concept and more or less definitive experiments which can be replicated, and which need little statistical analysis to be understood and to be used for further increments in knowledge.

In the process the statistician often finds it convenient, in establishing the presence of certain effects, to set up null hypotheses which can be definitively rejected. But it is rarely the case that the statistician or scientists can set up a sharply stated hypothesis which is broadly applicable and get evidence that establishes it definitely. Failure to reject such a hypothesis is rarely the same as establishing it.

The hypothesis that one can apply independence in calculating $P(M|I)$ carries a great deal of baggage, and if stated so sharply, will certainly be rejected with enough data. What is more relevant would be a measure of how much error the use of this hypothesis will introduce into the estimation of $P(M|I)$. We have some data on this issue for some cases, but we do not have as much basis to generalize as we would have if we confined attention to siblings or near relatives.

There is another difficulty with standard statistical practice. In the sequential process of refining hypotheses and experiments it is customary to use 95% or 99% confidence limits. One may be concerned with the meaning of a statement that one is 95% confident that

$P(M|I) = 10^{-12}$. While such a statement is not made, even the conservative procedures suggested by the NRC report depend on the use of 95% confidence bounds.

Matching on 4 or 5 loci should be more than enough evidence to establish G. That will become clearer as growing data banks determine the frequency with which unrelated individuals match on 2,3,4 and 5 loci. In the meantime, scientific proof, where science has a capital S, will not be sufficient to convince defense experts who refuse to accept a model applying independence. But before long the scientific literature will be pretty much in agreement on the strength of DNA evidence, just as it is on the claim that cigarette smoking causes lung cancer, while cigarette company executives feel free to doubt that in testimony before Congress.

12. DNA Evidence in Court

At first DNA evidence was accepted without question. As defense lawyers brought in expert witnesses, often with limited expertise, who questioned the quality and analysis, many courts excluded some or all aspects of the DNA evidence. While some courts insisted on quantitative evidence, others refuse to accept "statistical" evidence. One issue was the subjective nature of the decision to declare a match, which gave rise in some courts to a desire to automate that process.

The NRC report attempted to put the entire matter in perspective. As a result, enterprising lawyers found openings in some vague language and some of the criticism of the report. For example, Weir (1993), in testifying for the prosecution, rejected HW equilibrium at two loci on two of three population data sets tested. The defense lawyer convinced the judge that the NRC instructions implied that both data sets should be excluded and that the remaining data set did not suffice for NRC standards, and the DNA evidence should be inadmissible. A more reasonable interpretation would have been merely to eliminate the data at the involved loci-population pairs.

In reaction to the controversy about the NRC report, in disregard of the fact that many, if not most of the criticisms were directed toward the *excessive conservatism*, a number of courts chose not to admit DNA testimony (Kaye 1993, p. 148). Recently the decisions are beginning to move toward admitting DNA profile evidence and allowing the use of quantitative conclusions. The Minnesota Supreme Court decided to admit quantitative evidence after being faced with a threat by the state legislature to legislate its admissiblility (Zack 1994) and the New York Court of Appeals (Fisher 1994) made a decision in favor of admitting DNA evidence.

In the meantime the NRC is in the process of preparing another report, presumably to correct some minor errors and to clarify the points on which the previous report was vague.

13. Presenting the Evidence

A major argument against admissibility of quantitative evidence has been the fear that juries would not understand and would be unduly influenced by such numbers. It has been suggested that some judges automatically *tune out* when numbers are discussed and therefore prefer more qualitative statements.

If a number is demanded, then there is a problem with declaring a probability of 10^{-12} in a world with 5 billion people. My suspicion is that rather than being unduly influenced to believe in guilt, the jury is liable to be skeptical of the result or unsure of how to interpret it. Certainly the defense lawyer could easily counteract any undue influence by pointing out the weakness of the chain of evidence from G to guilt. In the meantime the prosecution has a problem of showing how such numbers can be derived from samples of a few hundred subjects. A table of random digits could be the source of a tutorial on how probabilities of 10^{-12} are easily attained and, while events of such low probabilities will occur, we should be surprised to see them repeated.

The principle of calculating posterior odds from prior odds and the likelihood ratio seems simple enough to me, now that I am an experienced statistician. Geisser (1990), a devoted Bayesian, is convinced that such analysis would not be appreciated by most juries and judges. Simple tables showing what proportion of siblings match on one, two, ... loci can help indicate forcefully to juries the strength of the evidence.

In summary, part of the function of the expert witness for the prosecution should be the design of a proper presentation of the evidence to inform the judge and jury of what the evidence does and does not mean.

14. Conclusions

There continues to be controversy about the force of evidence of matching DNA profiles. I predict that that controversy will diminish as courts realize that with proper quality control and real expertise, the current methods are ordinarily quite adequate to establish matching for legal purposes. It may take a few years for the experts to learn to deal with the few places where the NRC opened a door for clever defense lawyers.

It has been claimed that in a few years, we will have readily available direct sequencing methods which will describe the loci precisely without measurement error, and thereby eliminate the controversy. Those advances will be useful, but they will not dispose of the basic philosophical issues. The experts who argue that the numbers presented are meaningless because the models don't hold, will be able to argue that point just as forcefully in the future.

What is liable to be more convincing, now as well as in the future, is the point essentially made by Wooley and Harmon (1992). The prosecutor can ask "Why, if you think the suspect is innocent, you do not investigate another locus which is extremely likely to establish his innocence?"

¿From the point of view of when should new methods of scientific evidence be admitted, the legal profession is wise in not requiring unanimity among experts because such unanimity is difficult to deliver in fields as tenuous as science. On the other hand it is clear that the legal profession should move quickly to enhance the debate and analysis necessary to clarify the pros and cons of important new innovations. The debate tends to carry on long after the main issues are understood.

This paper has neglected the privacy and ethical issues in DNA profiling which are important and have been treated in the NRC report. I wish to thank D. Balding, J.L. Gastwirth, K. Lange, R.C. Lewontin, B.D. Spencer, and B.S. Weir for helpful discussions and D.H. Kaye for his article from which I borrowed much.

15. References

Aldhous, P. (1993). Geneticists attack NRC report as scientifically flawed. *Science.* **259**, 755-6.

Berry, D.A. (1991). Inferences using DNA profiling in forensic identification and paternity cases. (with discussion) *Statistical Science.* **6**, 175-205.

Cohen, J.E. (1992). The ceiling principle is not always conservative in assigning genotypes for forensic D testing. *American Journal of Human Genetics.* **51**, 1165-8.

Committee on DNA Technology in Forensic Inference (1992). *DNA technology in forensic science.* National Academy Press, Washington, D.C.

Devlin, B., Risch, N., and Roeder, K. (1990). No excess of homozygosity at loci used for DNA fingerprinting. *Science.* **249**, 1416-20

Devlin, B., Risch, N., and Roeder, K. (1993). Statistical evaluation of DNA fingerprinting: A critique of the NRC's report. *Science.* **259**, 748-

Fisher, I. (1994). Ruling allows DNA testing as evidence. *The New York Times.* Mar 30, B1

Gastwirth, J.L. (1988). *Statistical reasoning in law and public policy.* Vol. 2. Academic Press, Boston.

Geisser, S. (1990). Some remarks on DNA fingerprinting. *Chance: New directions for statistics and computing.* **3**, 8-9.

Jeffreys, A.J., Wilson,V., and Thein, S.L. (1985). Individual-specific 'fingerprints' of human DNA. *Nature.* **316** , 76-9.

Kaye, D.H. (1993). DNA evidence: Probablity, population genetics and the courts. *Harvard Journal of Law and Technology.* **7**, 101-72.

Kolata, G. (1991). Critic of genetic fingerprint testing tells of pressure to withdraw paper. *The New York Times.* Dec. 20, A16.

Lewontin, R.C. (1992). The dream of the human genome. *The New York Review,* May 28, 31-40.

Thompson, W.C. and Ford, S. (1989). DNA typing: Acceptance and weight of the new genetic identification tests. *Virginia Law Review.* **75**, 45-108.

Thompson, W.C. and Schumann, E.L. (1987). Interpretation of statistical evidence in court trials, The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior.* **11**, 167-87.

Wambaugh, J. (1989). *The Blooding.* Perigord Press, New York.

Weir, B.S. (1990). *Genetic data analysis: Methods for discrete population genetic data.* Sinnauer, Sunderland, MA

Weir, B.S. (1993). DNA fingerprinting report. *Science.* **260**, 473.

Weir, B.S. and Gaut, B.S. (1993). Matching and binning DNA fragments in forensic science. *Jurimetrics*). **34** 9-19 .

Weir, B.S. and Hill, W.G., (1993). Population genetics.(preprint). 1-18.

Wooley, J. and Harmon, R.P. (1992). The forensic DNA brouhaha: Science or debate. *American Journal of Human Genetics.* **51**, 1164-5.

Zack, M. (1994). Use of DNA evidence widened. *Star Tribune.* April 30, A1.

Abstract

The use, in court, of DNA Profiling, popularly referred to as DNA Fingerprinting, for forensic identification purposes has been questioned. A report of the National Research Council was solicited to clarify the issues and propose procedures of how and where this powerful technique could be used. This report has been subject to criticism. The main point of the report was to recommend procedures for the primary issue of quality control and for gathering useful data. The report also proposed the *ceiling principle* as a conservative approach for calculating the match probability for an innocent suspect to be used in assessing the guilt of a suspect. That method has been criticized by some as not necessarily conservative, and by others as unnecessarily conservative as well as illogical. These issues are discussed as well as some of the recent history in court.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Unlimited

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
TR No. ONR-C-16

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Dept. of Statistics
Harvard University

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS *(City, State, and ZIP Code)*
Department of Statistics SC713
Harvard University
Cambridge, MA 02138

7b. ADDRESS *(City, State, and ZIP Code)*

8a. NAME OF FUNDING / SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*
1111

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
N00014-91-J-1005

8c. ADDRESS *(City, State, and ZIP Code)*
Office of Naval Research
Arlington, VA 22217-5000

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO

11. TITLE *(Include Security Classification)*
Issues in DNA Fingerprinting

12. PERSONAL AUTHOR(S)
Professor Herman Chernoff

13a. TYPE OF REPORT
Technical

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Year, Month, Day)*
July 1994

15. PAGE COUNT
20

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

See reverse side.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED □ SAME AS RPT □ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION

22a. NAME OF RESPONSIBLE INDIVIDUAL
Herman Chernoff

22b. TELEPHONE *(Include Area Code)*
617-495-5462

22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED