Consider how we might change p_i and p_j in order to increase the value of this expression, however, keeping $p_i + p_j$ constant so that $\sum p_i = 1$. If the expression in square brackets is positive, the maximum occurs when $p_i = p_j$. If it is negative, then the maximum occurs when one of p_i and p_j is 0. Thus the expression is maximised when all the p_i are equal and $p_i = 1/n$.

Substituting these values into equation (1) gives

$$P^* = \frac{1}{n} \cdot \frac{n-1}{n} \cdots \frac{n-m}{n}. \tag{4}$$

It is simple to check that

$$\begin{aligned} &-\log\left(1-\frac{1}{n}\right) - \cdots - \log\left(1-\frac{m}{n}\right) \\ &\quad \geq \frac{m(m+1)}{2n}. \end{aligned} \tag{5}$$

Thus,

$$\max\ P^* \leq \frac{e^{-m(m+1)/2n}}{n}. \tag{6}$$

This is maximised by $n = m(m+1)/2$ giving

$$\max\ P^* \leq \frac{2e^{-1}}{m(m+1)}. \tag{7}$$

This is an appropriate p-value for the test of H: the accused is innocent. Even with a sample of size 100, the p-value is less than 1 in 10,000, and this seems sufficient for forensic purposes. It should be noted that under the alternative hypothesis H_1: the accused is guilty, max $P^* = 1$.

The analysis is not so simple when the number of matches in the sample is not 0, but the work is in progress. As is clear, the argument in no way depends on the categories being defined by DNA typing, but applies to any method of classification.

It is obviously of importance to choose a suitable criterion for a match. Usually these criteria have been based on some number of standard deviations of the error, without any stronger argument than that this should give a small probability of a mismatch. However, the most obvious course is to derive the criterion directly from a database. If this contains duplicate profiles, then it should be possible to devise a criterion which allows a very small percentage of false matches and a very high probability that two profiles from the same person will be declared a match. This has been shown to be possible by Herrin (1993) and Sudbury, Marinopoulos and Gunn (1993). A blanket criterion of allowing a 2 or 5% error for each band independently, neither takes into account band-shift nor the number of loci that have been successfully probed.

I enjoyed reading Kathryn Roeder's review of the DNA fingerprinting controversy and found it a fair-minded and comprehensive survey of the area. But has all this work really been necessary? Could we not have saved the courts a lot of trouble by keeping things simple?

Comment

William C. Thompson

To determine the value of forensic DNA evidence for proving two samples have a common source, one must take into account three sources of uncertainty. First, there is uncertainty about the interpretation of laboratory results. Were the bands in the DNA prints scored correctly? Has the analyst adequately accounted for any discrepancies between the "matching" prints? How likely are such discrepancies if the samples have a common source? Second, there is uncertainty about laboratory error. Could an error, such as inadvertant switching, mixing or cross-contamination of samples, have accounted for the incriminating results? How common are such errors? Third, there is uncertainty about the probability of a coincidental match. How rare are the matching genotypes?

Kathryn Roeder's review of the controversy over DNA fingerprinting focuses primarily on estimation of the frequency of matching genotypes. Her discussion of this intricate issue is helpful, although she might be faulted for failing to cite and discuss the arguments and data presented by other scholars who take a different point of view (e.g., Slimowitz and Cohen, 1993; Krane et al., 1992; Mueller, 1993; Geisser and Johnson, 1993). A more important complaint is that Roeder fails to take adequate account

William C. Thompson is Associate Professor of Criminology, Law and Society at University of California, Irvine, California 92717.

of the first two sources of uncertainty and that some of her comments about these issues are wrong or misleading.

UNCERTAIN MATCHES

In cases involving forensic RFLP analysis, there frequently is ambiguity in the scoring of bands (Thompson and Ford, 1991; Thompson, 1993). Most laboratories use computer-assisted imaging devices to score autorads, but the actual placement of bands and the ultimate determination of whether a band is present are within the analyst's discretion; manual overrides of the machines' scoring and placement of bands are common (Thompson, 1993). In some forensic cases, the scoring of a single ambiguous band can determine the outcome—one interpretation produces a damning incrimination of the defendant, another interpretation completely exonerates the defendant.

Ambiguity may also arise from discrepancies between DNA prints. In cases where a match is declared, the "match" often is imperfect because of some inconsistency in the number or position of bands (Thompson and Ford, 1991). When that happens, the analyst must judge whether the discrepancies reflect true genetic differences or are due to other factors, such as normal variation in the assays, degradation or mixing of samples or problems in the analysis, such as partial digestion, star activity or cross-hybridization. Such judgments also have a crucial bearing on the value of the evidence (Thompson and Ford, 1993).

NOTE 1. Thompson and Ford (1993) offer a Bayesian analysis of a criminal case in which a laboratory analyst and a defense expert disagreed about the interpretation of extra bands found in an evidentiary sample. Under the laboratory analyst's interpretation, the likelihood ratio describing the value of the evidence for incriminating the defendant was estimated to be 3 billion; under the defense expert's interpretation (which later proved accurate), the likelihood ratio was estimated to be less than 20.

The existence of interpretive ambiguity in RFLP analysis first became widely known as a result of the *Castro* case (Lander, 1989), but the sort of issues that arose in *Castro* are by no means unique or anomalous (Thompson and Ford, 1991; Thompson, 1993).

Given the pivotal role of expert judgment in cases with problematic matches, it is important to consider whether expert judgment in such cases is being exercised appropriately. Forensic laboratories have been criticized for making such judgments in a cavalier manner, without adequate scientific foundation (Lander, 1989, 1991; Shields, 1992; Thompson, 1993). Because laboratory analysts are not blinded to the identity of samples or the facts of the case, there is also a danger that such judgments will be contaminated by circular inference and logical bootstrapping. The analyst may infer that a discrepancy between two DNA profiles on one probe must be an artifact (rather than a true genetic difference) because there is a match on the other probes or, worse yet, because other evidence in the case suggests the two profiles have a common source. I heard one forensic analyst defend the scoring of an ambiguous band (a judgment that incriminated the defendant in a rape case) by saying "I must be right, they found the victim's purse in [the defendant's] apartment!" Such inferences are, of course, appropriately made by the trier-of-fact (e.g., the jury) in a criminal trial, not by an analyst who is providing a putatively objective, independent interpretation of the DNA test. This sort of bootstraping can convert problematic results into an apparently damning incrimination. In light of the potential seriousness of this problem, Roeder's defense of subjective criteria for matching seems painfully naive. [As Eric Lander (1989) put it: "When a result is reported to have an error rate of 1 in 100,000,000, it seems essential that the underlying data are not left as a matter of subjective opinion."] The best way to solve the problem would be to heed the National Research Council's (1992) call for objective standards for scoring and matching of bands (Thompson, 1993).

Under the current match/binning procedure, uncertainty about the match is not reflected in the statistics presented to the jury. The jury receives the same statistic (an estimate of the frequency of matching genotypes in a reference population) whether the interpretation was problematic or not, and is typically told that this statistic is an index of the value of the DNA evidence. When the "match" is problematic, the frequency of matching genotypes may have little or no relationship to the probative value of the DNA evidence and hence is at best an unhelpful statistic and at worse seriously misleading.

Roeder advocates a likelihood ratio approach (Berry, 1991; Evett, Scranage and Pinchin, 1993) in order to "obviate the need to declare a match" and thereby "avoid a great deal of argument in the courts." If I were confident that the methods used to compute the likelihood ratios accurately take into account the sort of uncertainties about interpretation discussed here, and that the resulting likelihood ratios can be communicated successfully to juries, I would support this proposal. Unfortunately, I am quite sure that the first condition has not been met and I am doubtful about the second. The likelihood ratio models discussed by Roeder (Devlin, Risch and Roeder, 1992; Berry, Evett and Pinchin, 1992) adjust the likelihood ratio in order to take into account discrepancies in the position and, in some instances, the number of bands in the prints being compared, but do not take into account uncertainty arising from such factors as the initial scoring of bands, the decision to ignore extra bands or the failure of laboratory controls. Adoption of these likelihood ratio approaches might well reduce courtroom arguments over matching, but it would do so by sweeping the underlying issue under the rug rather than addressing it fairly.

LABORATORY ERROR

In forensic laboratories, samples from a given case are typically processed together, in a batch, through a sequence of procedures that require manual transfers of genetic material from container to container. Inadvertant switching, mixing or cross-contamination of samples can cause false matches (Thompson and Ford, 1989). Although false matches do not always cause false incriminations, they have that potential. [One such error in a rape case caused an embarrassing but nonincriminating match between the defendant and the victim! (Thompson and Ford, 1991, page 143).] For example, an innocent suspect's DNA might inadvertantly be mixed with evidentiary samples during a sample transfer or gel loading, causing the suspect falsely to match a sample from the crime scene. Switching, mixing and cross-contamination of samples have been observed to occur in proficiency testing and actual forensic work with sufficient frequency to raise serious concerns (Thompson and Ford, 1991, pages 111–116; Koehler, 1993).

The rate of such errors is a function of the number of samples tested rather than (as Roeder would have it) the number of pairwise comparisons later made among the DNA prints of the samples. Hence, it is fallacious for Roeder to translate a laboratory's error rate on the California Association of Crime Laboratory Directors' (CACLD) proficiency test (one false match in each of two trials involving 50 samples) into an estimated error rate of 0.0008 for the laboratory. [Roeder fails to mention that the laboratory in question participated in two rounds of proficiency testing; in each round it was asked to compare 50 samples, and in each round it produced false match. For a detailed account of the study see Thompson and Ford (1991) or Koehler (1993).] However, it is also erroneous to equate the rate of false matches with the probability of a false incrimination because not all false matches would incriminate an innocent suspect. A better estimate of the probability of a false incrimination could be obtained by dividing the number of ways an incriminating false match could occur by the total number of ways a false match could occur in a given case, and then multiplying the quotient by the rate of false matches. In a case where a blood stain at a crime scene (B) is compared to DNA samples from the victim (V) and a suspect (S), for example, there are three possible false matches (S and V, S and B, V and B), only one of which would falsely incriminate an innocent suspect (S and B), so if the rate of false matches were 0.02, as suggested by the CACLD proficiency test, the probability of a false incrimination would be $0.02 \times \frac{1}{3} = 0.00666$.

The National Research Council (1992, page 88) has called for laboratory error rates to be determined based on proficiency testing and disclosed to juries. According to the NRC report, accurate estimates of error rate require proficiency tests that are externally administered, are blind and based on samples that are truly representative of case materials. Having an accurate estimate of laboratory error rates is probably far more important than having an accurate estimate of genotype frequencies because the former is likely to be much higher than the latter. To date, however, relatively little proficiency testing has been done, and most of it has not been blind (Koehler, 1993; Thompson and Ford, 1991). Hence, there currently is not a firm scientific basis for determining the rate of false positives.

Juries often hear nothing about false positives other than broad assurances that they never occur. [See *People v. Shi Fu Huang*, 546 N.Y.S.2d 920 (Co. Ct. 1989) ("Dr. Baird testified that it is impossible to get a false positive"); *People v. Wesley*, 533 N.Y.S.2d 643 (Co. Ct. 1988) ("it is impossible under the scientific principles, technology and procedures of DNA Fingerprinting (outside of an identical twin), to get a "false positive"—i.e., to identify the wrong individual as the contributor of the DNA being tested...Under the undisputed testimony received at the hearing, no 'wrong' person, within the established powers of identity for the test, can be identified."); *Hicks v. State* (Tex. Ct. Crim. App., No. 70,803, March 31, 1993) ("According to Caskey, a false positive finding was impossible..."); *State v. Cobey*, 559 A.2d 391, 392 (Md. App. 1989) ("An incorrect match is an impossible result"); also Koehler (1993) (quoting a number of similar statements from transcripts of expert testimony).] When such claims are challenged, forensic experts typically concede that false matches are possible but claim the likelihood of such an event is vanishingly small and that a false match has never occurred in their laboratory or, if it has, that it resulted from problems that have been corrected and will not reoccur. Experts for the defense sometimes testify that false positive are possible, but they typically do not attempt to quantify their frequency. (Thompson, 1993). As a result, jurors hear impressive numbers that appear to quantify with precision the frequency of the matching genotypes, accompanied (when the issue is raised at all) by a vague, nonquantitative discussion of the chances of a false positive.

The danger of the current approach is that the possibility of a false positive will simply be ignored. Indeed, the considerable time devoted in some trials to discussion of the frequency of the matching genotypes, the methods for computing that frequency and the controversy surrounding those methods may reinforce the powerfully prejudicial suggestion that false positives are a minor issue and that the frequency of the matching genotypes is the issue on which the value of DNA evidence will turn. In fact, where false positives are possible, the frequency of matching genotypes may have no relationship to the likelihood ratio that describes the value of the DNA evidence for proving two samples have a common source. Hence, it is at best an unhelpful statistic and at worse seriously misleading. Whether it should even be presented to juries is a question that I hope Kathryn Roeder, and her readers, will ponder.

Comment

B. S. Weir

Roeder has provided a useful review of the statistical issues involved in studies of human identification. She makes the distinction between objections to certain assumptions that might be raised in theory and the numerical consequences of those assumptions not being completely true in practice. A related issue is that of statisticians not taking into account all the relevant biological factors, and Roeder pointed to work of Geisser and Johnson (1992, 1993) in that context.

As Roeder explained, Geisser and Johnson explored the consequences of discretizing VNTR fragment lengths into a set of quantile bins, rather than the bins defined by viral fragment lengths as used by the FBI. Both binning strategies are adhoc, but the quantile bins lead to simpler analyses since each bin and each pair of bins is equally frequent. Roeder pointed out that the analyses of Geisser and Johnson have little relevance in the forensic debate since the problem of the unknown cause for single bands was ignored. The same point was made by Weir (1993), who also demonstrated that different numbers of bins, let alone different binning strategies, can lead to different conclusions regarding the independence of pairs of fragments in samples. The phenomenon has been well-documented in the population genetics literature.

Roeder herself might have referred to previous literature in her discussion of hierarchical Bayesian methods that invoke the Dirichlet distribution. Other authors have sought to use this distribution in the population genetics context (Rothman, Sing and Templeton, 1974; Spielman, Neel and Li, 1977), and there may be instances where it provides useful approximations. The current problem is to determine the conditional probability of a genotype, or VNTR profile, when that genotype has already been observed (for the perpetrator of a crime). Such conditional probabilities require the joint probabilities of *genotypes*, whereas Roeder in her equation (8) works with the joint probabilities of *alleles*. The joint genotypic frequencies require information about the relations between four alleles (two per genotype) rather than just two. Nichols and Balding (1991), in the paper that presented Roeder's equations (18), also ignored the relations between alleles considered three or four at a time. It is possible to approximate the necessary four-gene measures of identity with the two-gene measure called θ_s by Roeder, and θ or F_{ST} by others (Weir, 1994).

A deeper question concerns estimation procedures for θ. This quantity provides the correlation for alleles within the same subpopulation, and consequently it provides the component of variance between subpopulations in an analysis-of-variance setting. Evidently such a parameter cannot be estimated from data in one subpopulation (e.g., Weir and Cockerham, 1984), or even from data from the whole population without knowledge of subpopulation structure. Apparently, Roeder et al. (1993) overcome this logical barrier in arriving at estimates by assuming a distribution for allele frequencies, in contrast to the approach of Cockerham (1969) that regards the true allele frequencies as unknown.

The problem with taking genotypic frequencies to have a Dirichlet distribution is that results contrary to genetic expectations can result. Jiang and Cockerham (1987) simulated populations subject to genetic drift and compared a moment estimator of θ derived from an analysis-of-variance viewpoint with

B. S. Weir is in the Program in Statistical Genetics of the Department of Statistics at North Carolina State University, Raleigh, North Carolina 27695-8203, and has the title of William Neal Reynolds Professor of Statistics and Genetics. He directs an NIH-funded Program Project in Statistical and Quantitative Genetics that supports theoretical and experimental research in the Departments of Statistics and Genetics.