guilt, although it would be very strong evidence. If there were identical twins involved, or even siblings, the force of evidence might be reduced. Matching of band weights for several probes does not imply that the entire DNA system is the same. In fact, if only a few probes with few alleles matched, the evidence could be far from overwhelming, even in those cases where there is no error in measurement.

It should be remarked that if Lifecode type bins are used in a nonexclusionary fashion, apparent matching on almost all of the highly polymorphic probes could be strong evidence, even if there were apparent failure to match on one or two probes. The strength of the evidence would depend on the bin sizes and the relative magnitude of the errors of measurement.

Conditioning. One advantage of Bayesian analysis over NP analysis is that, in the former, we can examine the evidence as it arrives. A partial reply involves the use of conditioning. The force of classical inference is sometimes strengthened by using conditioning appropriately. That could be done here, too. For example, once we had measurements on Ponce's blood, we could use those data to help select what constitute effective probes for the comparison.

ACKNOWLEDGMENT

This research was supported in part by NSF Grant DMS-88-17204.

Comment: Uncertainty in DNA Profile Evidence

D. H. Kaye

Donald Berry's article on inferring identity from DNA profiles presents a method for "direct calculation of the probability that the suspect is guilty" and "the probability that an alleged father of a child is the true father." The method is Bayesian. Berry computes the posterior odds of guilt as the product of the prior odds (assessed on the basis of all the evidence apart from the electrophoretic measurements) and a likelihood ratio for the DNA results. The specific likelihood ratio (R) that he skillfully derives for single-locus restriction fragment length polymorphisms accounts for random measurement error in electrophoresis and for sampling error in a laboratory's data base on the distribution of fragment lengths in the population.

This comment examines, from the perspective of a lawyer, two connected issues: the forensic importance of quantifying measurement and sampling error and the desirability of combining likelihoods and priors for jurors or judges. I try to place Berry's treatment of these matters in the context of the emerging case law on DNA profiling, and I speculate about the advisability of bringing Bayes to the bar. Proceeding on the premise that Berry's mathematics is impeccable, I conclude that if "To bin or not to bin?" is the question, then Berry has the answer.

D. H. Kaye is Regents Professor, Arizona State University, College of Law, Tempe, Arizona 85287-7906.

1. LABORATORY MEASUREMENT ERROR

Berry's analysis handles normally distributed (and log normal) laboratory errors in measuring the position of a perceived band, and he notes that a more complex analysis could handle other continuous error distributions. Yet, much of the criticism of forensic DNA work emphasizes other threats, such as contamination and degradation of samples. Thompson and Ford (1991, page 138), for example, report that missing bands, extra bands and systematically shifted bands are "quite common in the forensic casework."

These types of experimental error have received considerable judicial attention. Virtually all courts in the United States to face the issue have held that DNA findings of identity are potentially admissible, but the degree of experimental rigor actually required for admission varies with the understanding of the court and the persuasiveness of the experts. *People v. Castro* (1989) is remarkable for the extent of the judicial inquiry into

methodology. After experiencing "a piercing attack upon each molecule of the evidence presented" generating 12 weeks and 5000 pages of expert testimony, the court concluded that Lifecodes "failed in its responsibility to perform the accepted scientific techniques and experiments in several major respects." Although the defense successfully challenged the inconsistency between Lifecodes' methods of declaring a match and its procedure for binning in the population data base, it apparently did not question the appropriateness of matching and binning per se. Instead, problems in interpreting missing bands and extra bands convinced the court to exclude Lifecodes' finding of a match.

The risks of missing and extra bands, as opposed to randomly misplaced bands, do not enter into the equations for the likelihood ratio R that is supposed to express the probative value of the DNA fragment measurements. Laboratory determinations of the jiggle—the standard deviation—in the measured locations of the same band derived from repeated electrophoreses in which the band that appears on each gel will not account for degradation and contamination. Although these experimental artifacts surely affect the probability of guilt conditioned on the DNA evidence, they play no part in the normal error model of laboratory measurements or obvious variations on it. Berry is well aware of these issues, and I discuss his treatment of them in Section 3.

In contrast to the issues of missing and extra bands, which well-litigated cases have exposed, the existence of random measurement error, which Berry forcefully addresses, seems to receive little judicial scrutiny. In focusing more attention on this problem, Berry's work should sharpen and improve the legal response to DNA profiling. As it stands, even when the defendant does pursue this line of attack, DNA testers have been known to deny the reality of problem. Thompson and Ford (1991, page 94, note 8) cite an Indiana case in which the following exchange occurred:

Q (Defense attorney): If we were to test two different samples from the same individual and come up with a DNA fingerprint, is there any variability in the test result in sample one and sample two? Might we expect there would be some variability in the two?

A (Prosecution expert): I wouldn't expect there to be any.

Q: They would be exactly the same?

A: Yep.

Somewhat more plausible is the testimony of FBI employees in *United States v. Jakobetz* (1990). In that case, the FBI declared a match whenever the bands being compared lay within a preset distance, it used wide bins to derive frequencies in its population data base and it ignored matching fragments of lengths not well represented in the population data base. While, in particular cases, this procedure could produce conservative estimates of $P(X \mid I)$—the likelihood of the DNA evidence under the hypotheses of innocence—Berry's approach seems more balanced and reasonable.

2. SAMPLING ERROR

The second source of error that affects R is sampling variability in the estimated distribution of fragment lengths in the relevant population. Berry demonstrates that even if there were no errors in the recorded positions of the bands in the samples of DNA used to construct population data bases, the instability of estimates in regions where the data are sparse can introduce serious uncertainty into inferences of identity.

Again, there are related threats to the inference of guilt that have yet to be subsumed in a single likelihood ratio. An issue of at least equal prevalence and difficulty as data base sampling error is the determination of the appropriate reference population. Thus, Berry observes that the basis for using "Hispanic" frequencies in *People v. Castro* (1989) was extremely weak, and he recommends looking at and perhaps averaging likelihoods for other racial categories. But there is concern that these categories are themselves too broad. It is quite possible, for instance, that frequencies of certain RFLP patterns in Mexican-Americans differ from citizens of Puerto Rican descent, which both could differ from individuals of Cuban or El Salvadoran heritage (OTA 1990, page 68). Likewise, in *United States v. Two Bulls* (1990, note 2), the FBI reported that the "probability of someone other than Two Bulls providing a match" to the semen stain on the underwear of a 14-year-old girl who was raped on a South Dakota Indian reservation "was one in 177,000, based on a Native American population base." The premise that Native Americans constitute a homogeneous genetic group is questionable. As Berry recognizes, population substructures and the lack of random assortment of genes across these subgroups could cast doubt on the computed value of R (but see *United States v. Jakobetz*, testimony of Kenneth Kidd that frequency differences for VNTRs between many subgroups are insubstantial).

In some cases, the courts have dealt with these threats to inferences of identity by reducing the match-binning probabilities. In *People v. Castro* (1989, page 993), the court suggested that

"[c]onservative or reduced calculations may also correct the Hardy–Weinberg deviation problems." Citing this dictum, the Georgia Supreme Court in *Caldwell v. State* (1990, page 444), approved of "a more conservative approach" that reduced $P(X \mid I)$ from 0.42×10^{-7} (obtained by multiplying individual band frequencies in Lifecodes' data base) to 0.40×10^{-5} (obtained by using the "data base itself, and not 'any population theory' "). Similarly to circumvent arguments about sampling variability in *United States v. Jakobetz* (1990, page 259), "the FBI declined to use the frequency data for two alleles on one probe because the alleles fell within [a] rare area of the gel."

(*Caldwell* is also of interest in that the defendant argued that Lifecodes should not have used "a double integral Gaussian weighted average" (presumably, the procedure of Morris, Sanda and Glassberg, 1989). To counter this complaint, Lifecodes used "a straight binning method" to find $P(X \mid I)$. 393 So.2d at 443. Berry (p. 179) characterizes the Morris procedure as "roughly equivalent to using a bin size of $\pm 2/3$ s.d.'s.")

These "solutions" are hardly ideal. The approach Berry advances seems a better response to the problem of sampling error in data bases. With further refinement it may be capable of handling even correlated errors due to band shifting. Nonetheless, it is not clear to me how "extending the likelihood ratio to handle the dependent case" (page 4) will cope with uncertainty or bias arising from population substructure. This possible limitation in the analysis has implications for the presentation of likelihoods in court, and it is to this topic that I now turn.

3. BAYESIAN INFERENCE IN COURT

3.1 The Pure Opinion Format

Should forensic DNA laboratories use probabilities or odds to express the uncertainty in their findings, and, if so, which probabilities should they report? One can imagine a world in which numbers are verboten, and experts are constrained to stating categorical opinions. The report from Cellmark Diagnostic Corporation in *State v. Schwartz* (1989, page 424), for example, concluded that "it is the opinion of the undersigned that the DNA banding patterns obtained from the stain removed from the blue jeans and the blood of Carrie Coonrod are from the same individual." Likewise, experts from Lifecodes testified in *Martinez v. State* (1989, page 695) "that the DNA from the crime semen sample and from Martinez's blood sample were taken from the same individual." The difficulty with the pure opinion format is, of course, that jurors lack adequate information to decide how much credence they should give to such opinions.

Consequently, although lawyers usually push for opinions expressed in no uncertain terms, the law does not force experts to give them, and no jurisdiction excludes all well-founded probability calculations from all cases. The state with the broadest exclusionary rule is, as Berry notes, Minnesota. Beginning with *State v. Carlson* (1978) and culminating most recently in *State v. Schwartz* (1989), the Supreme Court of Minnesota, relying on certain arguments in Tribe (1971) against the kind of Bayesian presentation that Berry favors, held a wide variety of probabilities and population frequencies inadmissible in criminal cases. Not all such statistics are inadmissible, however. *State v. Kim* (1987) holds that the relative frequency of each independent genetic marker is admissible; only the product of these frequencies is not. This rule makes little sense, and it has been overturned by the legislature, at least as to probabilities involving genetic markers (Minnesota Statutes §634.26 (Supplement 1989)).

3.2 The Improbability Format

Instead of adhering to the pure opinion format, courts allow probabilities or relative frequencies that could help the jury come to a decision. These numbers typically accompany the expert's opinion. In *Schwartz*, Cellmark reported that the frequency of this DNA banding pattern in the Caucasian population is approximately 1 in 33 billion" (Cellmark used multilocus probes. The logic that generates the 1 in 33 billion figure is detailed in Kaye 1990b, and is essentially identical to that used in *Robinson v. Mandell*, 1868, as described in Meier and Zabell, 1980). In *Martinez*, Lifecodes' laboratory supervisor testified as follows:

Q: And what would be the answer to that question as far as the likelihood of finding another individual whose bands would match-up in the same fashion as this?

A: The final number was that you would expect to find only one individual in 234 billion that would have the same banding that we found in this case.

Q: What is the total earth population, if you know?

A: Five billion.

Q: That is in excess of the number of people today?

A: Yes. Basically that's what that number ultimately means is that that pattern is unique within the population of this planet.

Like Berry, I find this kind of testimony disquieting. Nevertheless, the matching and binning probability is informative. Due to the problems of measurement error, sampling error, and contamination, degradation, population misidentification and substructure, and the lack of independence in band positions, the number testified to is not $P(X \mid I)$, the conditional probability of a reported match under the hypothesis of innocence. Instead, it is the conditional probability of a true match in the case of an innocent suspect and a proper (and error-free) population data base. Knowing this number is better than nothing. Therefore, one might argue, the legal system should allow the jury to see the number and leave it to the defendant to point out its limitations and demonstrate that the likelihood for innocence is much smaller than a match-binning probability like 1/234 billion.

This view might have some merit if all criminal cases went to trial and all defendants had access to astute and skilled defense attorneys. In fact, the vast majority of cases never reach trial and very few defense lawyers have the knowledge and resources required to expose the limitations and imperfections in the match-binning probability as a measure of innocence. Prosecutions will be instituted, and most defendants will plead guilty in the face of such astronomical numbers. Since there is every reason to get things right before trial, Berry's efforts to incorporate more of the sources of error into a reported likelihood ratio are appealing.

3.3 The Bayesian Format

It does not follow, however, that a full-blown Bayesian presentation is advisable. The appropriateness of Bayesian methods, particularly in criminal cases, has been the subject of a painfully prolonged—and sometimes muddled—debate among legal scholars (Kaye 1988a). Partly because Bayesian methods rarely are employed in the legal system, the courts have had less to say about the matter. In 1979, Ellman and Kaye pointed to the practice of introducing posterior probabilities (computed under the ad hoc and often undisclosed selection of a prior probability of 1/2) in civil paternity cases involving testing of antigens. Yet, they suggested that the courts did not recognize the Bayesian nature of this paternity probability. Subsequent legislative and judicial acceptance of these paternity probabilities is chronicled in Kaye (1990a, b; 1989; 1988b) and Kaye and Kanwischer (1988). Outside of the civil paternity area, Bayesian methods have not gained much of a foothold. A posterior probability computed by Lenth (1986) was prominent in a chain saw murder case, *State v. Klindt* (1986), but it is doubtful that the Iowa courts appreciated the basis of the calculation. Rape cases in which the prosecution relies on a "probability of paternity" have generated appellate opinions critical of that probability, but the criticisms relate to the arbitrariness of the prior odds. In short, no court has produced a thoughtful opinion as to the admissibility of reasonably produced posterior probabilities. The question must be considered open from the legal standpoint.

The argument for introducing Bayes's rule in court strikes me as less compelling than the case for using Berry's R. In this regard, it is helpful to distinguish between two ways of implementing the rule. In a "weak" Bayesian format advocated in, for instance, Ellman and Kaye (1979), the statistician merely explains how the likelihood ratio affects priors. He or she does not ask jurors to commit themselves to a particular prior or to use Bayes's rule to deduce a particular posterior. The theorem acts as a heuristic device, displaying the force of the evidence across a wide range of conceivable priors.

Berry, on the other hand, proposes a "strong" Bayesian format in which the statistician tackles the "more challenging problem" of helping jurors assess priors. He recommends "[c]omparing preferences for the prospect G as compared with well-understood bets (coins and dice)." Although many in the legal world would be scandalized by the idea of asking jurors to compare their preference for a bet on the defendant's guilt to a bet on the roll of the dice (see Cohen, 1981; Nesson, 1985; Tribe, 1971), I do not think that the legal system inevitably will compromise the basic human values if it asks jurors to return a verdict with a conscious and explicit awareness of the risk of a false conviction (see Shaviro, 1989).

Still, I do have two reservations about the strong Bayesian exercise. First, I wonder how well the typical juror will assess prior odds, even with some coaching about coins and dice. Admittedly, if R is astronomical, this may not matter. But if R is so huge, why bother with the Bayesian explanation? Why not just let the expert report that it is immensely more probable that the DNA samples would produce the observed bands if the two samples had a common source than if they came from different people? To the extent that the effort to teach numerically unschooled jurors how to find their prior odds does not seem worth the expected improvement in inferential accuracy, so to speak, the strong Bayesian format may not be justified. On the other hand, this argument carries less weight as applied to the weak Bayesian presentation. That approach does not require the juror to pick a specific prior probability but still can

convey some sense of the power of the laboratory's evidence.

Second, I worry that forcing jurors to articulate prior odds conditioned on the non-DNA evidence and to multiply this prior by a likelihood ratio may omit (and possibly divert attention from) major uncertainties in the experimental evidence. As indicated in Sections 1 and 2, the likelihood ratio R does not account for the risks of missing bands, extra bands, population misspecification and substructure.

One can respond, as I suspect Berry might, that all conceivable sources of error need not be reflected in a single figure for the posterior odds. One might treat Berry's analysis as conditioned on the absence of experimental embarassments, at least where the laboratory has observed rigorous protocols (compare OTA, 1990). Where it is not clear whether a suspect is homozygous or a fragment has gone undetected, one can compute distinct values of R under each assumption—as in Berry's discussion of *Castro*. Similarly, one can perform multiple computations of R and hence $P(G \mid X)$ for different racial categories.

The final result, however, is no longer a simple posterior probability for guilt or even a single table of posteriors and priors. It is a set of competing numbers or tables—accompanied, quite possibly, by some nagging doubts that must be left out of the equations for want of adequate data or analytic tools. If the residual uncertainty is substantial, then the jury must attend to it in some intuitive fashion anyway. It cannot take $P(G \mid X)$ at face value if the defendant (or the prosecution in a case in which the defendant offers an exculpatory DNA profile) raises serious questions about population structure or other uncertainties not included in sensitivity analysis of R and $P(G)$. And if this situation does materialize, one is left too wonder once again whether the expected payoff from the Bayesian format is worth the demands it places on the experts, the parties and the court.

4. CONCLUSION

As a lawyer, I see in Berry's article a cogent and powerful indictment of the matching and binning reasoning now used in single-locus DNA profiling. Berry builds an impressive case for using likelihoods that (a) make better use of the information in the test results and the population data and that (b) handle more of the uncertainties now present in DNA evidence.

I am less enamored of the strong Bayesian demand that jurors should quantify their prior probabilities and combine them with likelihood ratios based on certain simplifying assumptions to return a verdict of guilt or innocence. Like the courts, however, I am not prepared to say that there is no room for some form of a Bayesian presentation in a criminal trial. Considering the difficulties that many courts, attorneys and jurors face in assessing quantitative evidence, the efforts of Berry and other statisticians (e.g., Kadane, 1990; Fienberg and Kadane, 1983) to develop suitable Bayesian analyses for forensic applications are a most welcome development.

Comment

Ian Evett

Professor Berry's analysis of DNA profiling is elegant and penetrating. I will not discuss the detail of his treatment but will concentrate on issues touched on by the other discussants that are relevant to the work of the forensic scientist.

Ian Evett is on the staff at Central Research and Support Establishment, Home Office Forensic Science Service, Aldermaston, Reading, Berkshire, RG7 4PN, England.

First, should the forensic scientist adopt a Bayesian view of evidence evaluation? It has been the convention, from the first glimmerings of the science, to view evidence from a frequentist perspective. Consider a simple case where the evidence consists solely of a blood stain at a crime scene and there is a single suspect who gives a sample of blood. Assuming a system of discrete alleles with no measurement error then, if the suspect's blood and the scene blood are the same type—say X1—the scientist will refer to a data collection of some sort and, as well as reporting a