Are DNA Profiles Unique?

B.S. Weir
Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh NC 27695-8203.

INTRODUCTION

The presentation of DNA profile evidence in court has almost always been accompanied by some numerical statement of the form "The probability of finding this profile in a random person from the US Caucasian population is 1 in a million." As DNA profiles have become based on larger numbers of loci, the numbers have become more extreme and it is not unusual to see calculated values involving billions, trillions or beyond. There is a danger that these numbers will be counterproductive in the sense that it may appear difficult to assign them any credibility simply because they are so extreme.

Maybe because of this difficulty, the FBI announced at the Eighth International Symposium, and then at a Press Conference (reported in Science 278:1407, 1997) that "If the estimated probability of a DNA profile found in a crime sample is less than 1 in 260 billion, and it is seen in a person, then that person is the source of the sample." In such cases the numbers would not be reported. Indeed, in FBI Report 29D-OIC-LR-35063, the following statement was made "Based on the results of these seven genetic loci, specimen K39 (Clinton) is the source of the DNA obtained from specimen Q3243-1, to a reasonable degree of scientific accuracy."

In this note I review some of the history surrounding the question of uniqueness, and examine the basis for the FBI policy. There is a quite substantial scientific literature that considers the question of uniqueness, from the perspectives of forensic science, probability, and statistical genetics. Each will be considered in turn.

FORENSIC SCIENCE APPROACH

In the classic forensic science textbook "Crime Investigation" Kirk (1974) said that "The central problem of the criminal investigator is the establishment of personal identity - usually of the criminal, sometimes of the victim." He made a distinction between identity, meaning a unique existence, and individualization, pointing to a specific person. Kirk went on to say: "No two objects can ever be identical. They can and often do have properties that are not distinguishable. If enough of these properties exist ... identity of source is established." No two different things can be identical, and the DNA profiles from a suspect and a crime scene are different things. A fingerprint from a crime scene is not identical to a suspect's recorded fingerprint, but can be used to identify him and prove his individuality. In a prescient passage, Kirk further said "The criminalist of the future may well be able to individualize the criminal directly through the hair he has dropped, the blood he has shed, or the semen he has deposited. All these things are unique to the individual, just as his fingerprints are unique to him."

Therefore, the forensic science question is not "Is this profile unique?" (it is), and not "Are these two profiles identical?" (they can't be), but " Is there sufficient evidence to demonstrate that these two profiles originate from the same source?" In common usage a categorical statement of identity of source is called identification.

FINGERPRINTS

In spite of Galton's statistical calculations in the 1890s, probability arguments for the rarity of fingerprints were used very little at that time, and are not used at all now. In 1939 FBI Director Hoover wrote that fingerprints were "a certain and quick means of identification."

In a review of the history of fingerprints, Stigler (1995) suggested that the acceptance of uniqueness probably followed from "(i) a striking visual appearance of fingerprints in court, (ii) a few dramatically successful cases, and (iii) a long period in which they were used without a single case being noted where two different individuals exhibited the same pattern." Stigler anticipated the same growing acceptance of DNA profiles being unique.

PROBABILITY APPROACH

Statisticians have considered the role of probability theory in interpreting evidence. Mode (1963) anticipated the use of the product rule for DNA evidence in saying "Mathematical probability is the basis of much evidence presented in the court room, although it may not be recognized as such by lawyers and jurors. A number of individual circumstances, although singly of low evidential value, might jointly lead to but one conclusion. This has as its mathematical basis the law of compound probability for the occurrence of independent events."

Kingston (1965) defined partial transfer evidence as physical material or impressions transferred from crime scene to perpetrator (or perpetrator's possessions), or vice versa. He considered that such evidence can be characterized and assigned to an identity-set, leading to the questions "Does a particular person (or their type) belong to the set? Does anyone else belong to the set? " It is interesting to note that Mode, writing in a statistical journal, would include the following statement: "If it is highly improbable that another member could be found, we would be reasonably sure that the correct origin has been located. But if it is quite probable that other members exist, we would not be so sure that we have the correct origin.

Mode's language was echoed by the second NRC report (National Research Council, 1996): "The profile might be said to be unique if it is so rare that it becomes unreasonable to suppose that a second person in the population might have the same profile."

Mode's arguments can be formalized. Suppose p(x) is probability that an identity set has x members. Suppose also that P is the (estimated) probability that a random individual will belong to the set. If individuals are independent, in a population of size N, p(x) will be binomial B(N, P). Following the evidence of at least one member of the identity set, the probability of there being no other member of the set is

$$\Pr(x = 1 \mid 1 \le x \le N) = \frac{p(1)}{\sum_{x=1}^{N} p(x)}$$

and, when P=1/N, this is 0.58. If P=1/(1,000N), it is 0.9995.

The second NRC report (National Research Council, 1996) did not condition on the evidence type being seen once already, and calculated the probability of there being no-one in the population (of size N-1) being in the identity set. This probability is

$$\Pr(x = 0) = (1 - P)^{(N-1)}$$

and, if P=1/N, this is 0.37. If P=1/(1,000N), the probability is 0.999.

Evidently, there is not much difference between zero occurrences of a type in (N-1) individuals and one occurrence in N individuals given that it has been seen once provided independence of types within the population is assumed. Independence of profiles is key to the FBI policy. For a population of size N=260 million, when P=1/(1,000N)=1 in 260 billion, there is a 99.9% probability of a correct determination of source of the crime sample.

THE ISLAND PROBLEM

It is helpful at this point to refer to a classical (if artificial) forensic problem (Balding and Donnelly, 1995; Dawid and Mortera, 1996; Eggleston, 1983). Suppose a crime is committed on an island where N people live. A bloodstain was left by the perpetrator. One person is suspected, and then typed. He is found to match. What is the probability he left the bloodstain? Calculation of probabilities requires Bayes' theorem. If E is the evidence of a match between person and stain, H_p is the proposition that he left the stain, and H_d is the proposition that he did not:

$$\frac{\Pr(H_p \mid E)}{\Pr(H_d \mid E)} = \frac{\Pr(E \mid H_p)}{\Pr(E \mid H_d)} \times \frac{\Pr(H_p)}{\Pr(H_d)}$$
$$= \frac{1}{P} \times \frac{\Pr(H_p)}{\Pr(H_d)}$$

where P is the probability a random person would match. If equal priors are assumed, $\Pr(H_p)=1/N$. If P=1/N, then the posterior probability is $\Pr(H_p|E) = N/(2N-1) \cong 1/2$. At first sight it might appear surprising that the probability of a correct identification of the perpetrator is only one half, instead of one. The use of likelihood ratios is crucial.

LIKELIHOOD RATIOS

The use of likelihood ratios to weigh alternative propositions for DNA evidence has now been described in several textbooks: Aitken (1995), Evett and Weir (1998), Robertson and Vignaux (1995), Royall (1997) and Schumm (1994). An excellent summary was provided by Friedman (1996).

It appears as though courts are also beginning to see the logic of this approach. In the case Johannes Pruijsen v. H.M. Customs and Excise, the Crown Court in Chelmsford, United Kingdom said on July 30, 1998 "We note and we follow and accept unreservedly Dr. Evett's evidence to us and his strictures to us that we cannot look at one hypothesis, we must look at two and we must test each against the other... what is the probability of the evidence if the Respondent's hypothesis is correct? What is the probability of the evidence if the Appellant's hypothesis is correct? [Dr Evett] tells us (and we follow it) that if the answer to the first question is greater than the answer to the second question, then the Respondent's hypothesis is supported by the evidence."

The use of likelihood ratios should be distinguished from a Bayesian approach, which requires the specification of prior probabilities. Bayesian methods are unlikely to be used widely in forensic science any time soon. However, there has been a substantial increase in the number of Bayesian papers in the statistical literature, due in part to new computational techniques.

STATISTICAL GENETICS APPROACH

DNA profiles are genetic. They come with a structure and a history, and should be interpreted accordingly. In particular, notice should be given to the dependence among profiles due to family relationships, and to a shared evolutionary history.

EFFECTS OF RELATIVES

If suspect has a certain profile, the probability that his relative has the same profile is greater than the profile probability. At one locus, the reciprocal of these probabilities (when all allele frequencies are 0.1) are

Relationship	Homozygote & Heterozygote	
Full brothers	3.3	3.3
Father and son	10.0	10.0
Half brothers	18.2	16.7
Uncle and nephew	18.2	16.7
First cousins	30.8	25.0
Unrelated	100.0	50.0

It would be a mistake to ignore the effects of relatives when calculating the probability that some person other than a defendant had a particular DNA profile, unless there were good reasons to exclude all relatives.

In a related issue, Donnelly (1994) considered matching probabilities at different loci. If M_i indicates a match at locus i, the laws of probability give

$$\Pr(M_1M_2M_3\ldots) = \Pr(M_1)\Pr(M_2|M_1)\Pr(M_3|M_2M_1)\ldots$$

where the " | " sign indicates "conditional on." Only if the loci are independent does

$$\Pr(M_1M_2M_3\ldots) = \Pr(M_1)\Pr(M_2)\Pr(M_3)\ldots$$

Donnelly pointed out that, as more and more loci match, the more likely it is that people are related (if they are not the same person), and so the more likely it is that the next locus will also match. In other words, matching is then not independent over loci.

EVOLUTIONARY RELATEDNESS

If the history of the population has resulted in an average relationship \theta among pairs of alleles, the conditional probability of one homozygote AA given another is

$$\Pr(AA \mid AA) = \frac{[(1-\theta)\,pA + 2\theta][(1-\theta)\,pA + 3\theta]}{(1+\theta)(1+2\theta)}$$

and this is greater than the homozygote probability

$$\Pr(AA) = p_A^2 + \theta p_A(1-p_A)$$

In other words, having seen the profile once, it is more likely that it will be seen again. This is ignored in the second NRC report.

Moreover, if evolution is imposing a dependence between alleles at one locus, it is also imposing a dependence between alleles at different loci. There is a two-locus analog Θ of θ that allows an expression for two-locus genotypic frequencies. For individual homozygous for alleles A and B at two loci:

$$\Pr(AABB) = \Pr(AA)\Pr(BB) + (\Theta - \theta^2)\,p_A(1-p_A)\,p_B(1-p_B)$$

and this is greater than the product Pr(AA)Pr(BB).

CONCLUSION

It is very difficult to arrive at a satisfactory probabilistic or statistical genetic theory which will give the probability that a second person in a population has the same DNA profile as the one featuring in a criminal trial. The difficulties stem from possible dependencies between loci and between individuals. This brings us back to the situation described by Stigler for fingerprints, and led the 1996 NRC report to state: "The definition of uniqueness is outside our province. It is for the courts to decide … "

Uniqueness is not an issue that can be addressed with statistics.

REFERENCES

1. Aitken C.G.C. (1995) Statistics and the Evaluation of Evidence for Forensic Scientists. Wiley, New York.
2. Balding D., Donnelly P. (1995) Inference in Forensic Identification. J Roy Stat Soc (A), 158:21.
3. Dawid A.P., Mortera J. (1996) Coherent Analysis of Forensic Identification Evidence. J Roy Stat Soc (B), 58:425.

4. Eggleston R. (1983) Evidence, Proof and Probability, Weidenfeld & Nicholson, London.
5. Evett I.W., Weir B.S. (1998) Interpreting DNA Evidence. Sinauer, Sunderland, MA.
6. Friedman R.D. (1996) Assessing Evidence. Michigan Law Review, 94:1810-1838.
7. Kingston C.R. (1965) Applications of Probability Theory in Criminalistics. J Am Stat Assoc, 60:70-80.
8. Mode E.B. (1963) Probability and Criminalistics. J Am Stat Assoc, 58:628-640.
9. National Research Council. (1996) The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, D.C.
10. Robertson B., Vignaux G.A. (1995) Interpreting Evidence: Evaluating Forensic Science in the Courtroom, Wiley, New York.
11. Royall R. (1997) A Likelihood Theory of Evidence, Chapman and Hall, London.
12. Schumm D.A. (1994) Evidential Foundations of Probabilistic Reasoning, Wiley, New York.
13. Stigler S.M. (1995) Galton and Identification by Fingerprints. Genetics, 140:857-860. Program in Statistical Genetics Phone: (919) 515-3574 Department of Statistics Fax: (919) 515-7315 North Carolina State University URL: www.stat.ncsu.edu Raleigh NC 27695-8203 (click on “statistical genetics”)