quence among the "repeats" in the VNTR loci, so that each person can be recognized by a unique signature. If our interest is, indeed, to correctly identify the perpetrators of violent crimes, then it is unclear why we continue to argue about probability calculations and statistical artefacts in place of carrying out the necessary research to create a real "DNA fingerprinting."

Comment

Aidan Sudbury

Imagine a scientific world in which there is no theory of population genetics, but in which a clever technique has been devised which associates with each individual a set of six characteristics. Let us give this technique a name, say, "DNA fingerprinting." It is claimed that these fingerprints identify someone with a high degree of reliability. To test this viewpoint, databases are assembled and it is found that matches are indeed very rare—in fact, that a match between different individuals is found on average 1 in 10,000 times. Compared to other evidence accepted by the courts, such as identity parades, alibis, motives for the crime, this is considered very reliable and has become accepted.

Now, some years later, the theory of population genetics evolves, and a new method of determining match probabilities is based on this theory. Two things may happen. First, the calculations suggest match probabilities of the order of 1 in 10,000. In which case we may say "How interesting! But I don't think we want to burden the courts with the considerable complexities involved with these calculations. We're quite happy with the way we're doing things." Second, the calculations may suggest probabilities of the order of 1 in 100,000, in which case we shall just assume they are wrong.

To return to the real world: it seems that the undoubted charms of population genetics, with its Hardy–Weinberg and linkage equilibrium, have led us into confusing the primary with the secondary evidence. If the observed match probabilities in databases were not small, no amount of testing of databases for independence, or discussion as to just how different allele frequencies are in different races could persuade us that the theory we were using was correct.

As far as I know, all investigations of databases (see, e.g., Risch and Devlin, 1992a, b; Herrin, 1993; Sudbury, Marinopoulos and Gunn, 1993) have shown that matches between unrelated individuals are extremely unlikely. Among related individuals, only the immediate family (brother, sister, father, mother) are sufficiently close to give a probability of a match that is not forensically significant. What perhaps remains to be shown is that matches within small communities are still rare even though there has been a degree of inbreeding in the past. Nichols and Balding (1991) have treated this problem theoretically, but some data covering these situations would be welcome.

Now, let us see how knowledge about the number of matches in a database may be used to make statements about the probability of guilt. Suppose the population can be classified into an unknown number of categories $C_1, \ldots, C_n$ and that these have unknown frequencies $p_1, \ldots, p_n$, $\sum p_i = 1$. Further, a sample of size m has been taken and none have been found to be from the same category (there have been no matches). Now, a sample taken from the accused has been found to be in the same category as a crime sample, but both are different from any in the original sample. The aim is to use this data to test the hypothesis H: the accused is innocent.

The probability that the crime samples should match, but no others, under H is

$$(1) \qquad P^* = \sum_{i=1}^{n} p_i^2 \sum_{j_l \neq j_k \neq i} p_{j_1} \cdots p_{j_m}.$$

An appropriate p-value of the test is the maximum of this expression over all sets $\{p_i\}$. Consider the terms involving p_i and p_j. They are of the form

$$(2) \qquad A(p_i^2 p_j + p_j^2 p_i) + B(p_i^2 + p_j^2) + C p_i p_j + D(p_i + p_j),$$

where A, B, C and D are functions of the other p_l. This expression can be written

$$(3) \qquad \begin{aligned}&[A(p_i + p_j) + C - 2B]\, p_i p_j + B(p_i + p_j)^2\\ &\quad + D(p_i + p_j).\end{aligned}$$

Aidan Sudbury received a Ph.D. in Astrophysics from Monash University, where he is now a Senior Lecturer. His address is: Department of Mathematics, Monash University, Clayton 3168, Australia.