BIOMETRICS 55, 671–677
September 1999

Likelihood Ratios for Evaluating DNA Evidence When the Suspect is Found Through a Database Search

Anders Stockmarr

Department of Forensic Genetics, Institute of Forensic Medicine, and Department of Biostatistics, University of Copenhagen, 11 Frederik V's vej, DK-2100 Copenhagen, Denmark
email: anders.stockmarr@forensic.ku.dk

SUMMARY. A crime has been committed, and a DNA profile of the perpetrator is obtained from the crime scene. A suspect with a matching profile is found. The problem of evaluating this DNA evidence in a forensic context, when the suspect is found through a database search, is analysed through a likelihood approach. The recommendations of the National Research Council of the U.S. are derived in this setting as the proper way of evaluating the evidence when finiteness of the population of possible perpetrators is not taken into account. When a finite population of possible perpetrators may be assumed, it is possible to take account of the sampling process that resulted in the actual database, so one can deal with the problem where a large proportion of the possible perpetrators belongs to the database in question. It is shown that the last approach does not in general result in a greater weight being assigned to the evidence, though it does when a sufficiently large amount of the possible perpetrators are in the database. The value of the likelihood ratio corresponding to the probable cause setting constitutes an upper bound for this weight, and the upper bound is only attained when all but one of the possible perpetrators are in the database.

KEY WORDS: Database searches; DNA profiling; Forensic genetics; Likelihood ratio; One over n rule.

1. Introduction

DNA profiling techniques have become frequently used methods in forensic identification. If a crime has been committed and the perpetrator has left traces at the crime scene, e.g., blood, semen, or tissue, a DNA profile may be obtained. A suspected perpetrator's DNA profile is then compared to the one from the true perpetrator, and discordance in the comparison will exonerate the suspect while a match between the two profiles will tend to strengthen the belief that the suspect is actually the true perpetrator. A need for guidelines on the evaluation of the results of the use of these techniques resulted in the report "*DNA Technology in Forensic Science*" from the National Research Council on DNA Forensic Science (NRC) of the U.S. in 1992. In 1996, after much criticism of the recommendations in the 1992 report on how to evaluate the DNA evidence statistically (e.g., Balding and Nichols, 1994; Roeder, 1994; Balding and Donnelly, 1995a; Aitken, 1995), the NRC published a second report, "*An Update: The Evaluation of Forensic DNA Evidence*" (from here on, denoted as the NRC II report) with revised recommendations for the use of DNA technology in forensic science. This (and the former) report distinguishes between two cases of identification, namely the case of probable cause (Balding and Donnelly, 1995b), where the suspect has been identified on grounds that are not related to his DNA profile, and the case where the suspect is identified through a database search, i.e., he/she is identified because a DNA profile in the database happened to match the one observed at the crime scene. In the latter case, a much lower weight in favour of the suspect being the true perpetrator should be assigned to the evidence compared to the former case. The reason for this is the possibility of observing a match by chance when searching the database. This distinction and the decrease in the weight of the evidence have been subject to much criticism in the forensic community at large and in the literature. It has been claimed, mainly by Balding and Donnelly (1995b, 1996), and cited by several authors, e.g., Robertson and Vigneaux (1995), Weir (1996b), Lempert (1997), Taroni and Aitken (1997), and again by Balding (1996, 1997) and recently in the book by Evett and Weir (1998), that the weight of the evidence against a suspect who is identified through a database search should not be less than in the probable cause case but actually slightly larger since there is information in the fact that all the persons in the database but the suspect have been excluded. Furthermore, it has been argued that the recommendation given in the NRC II report is in conflict with the generally accepted principle that the DNA evidence should be evaluated through a likelihood ratio. These claims may be very serious to persons suspected of a crime and taken to trial because of identification through a database search, and they are virtually unopposed in the literature when the point of view is taken that the DNA evidence should be evaluated statistically through a likelihood ratio. One exception is Morton (1997). But a close examination of what the experiment 'searching a database' actually is shows that none of these claims hold, and the purpose of this paper is to place the recommendation of the NRC II report

in the case of identification through a database search in the proper probabilistic framework, as the likelihood ratio associated with a pair of competing hypotheses, and to argue that only this pair of hypotheses should be considered. As a consequence, it will be argued that the route taken by Balding and Donnelly (1996) is inappropriate.

2. Likelihood Ratios

Suppose that a crime has been committed and that a DNA profile sample $\mathcal{E}_{\mathrm{TP}}$ is obtained from material left at the crime scene. Use TP to denote the donor of the sample, as an abbreviation of the 'true perpetrator' for convenience, although we are not discussing whether this person actually committed the crime. The true perpetrator is considered as a member of a population of possible perpetrators, which may or may not be of infinite size, although it should not usually be identical to, e.g., the persons living in a given city. In a rape case where $\mathcal{E}_{\mathrm{TP}}$ is obtained from a semen stain left at the scene, no women should be included in the population of possible perpetrators. Special problems that arise in the case of a finite population of possible perpetrators will be dealt with in Section 3. The DNA profiles of the possible perpetrators will be referred to as the population of profiles, $\mathcal{P}$, and by a database $\mathcal{D}$ I shall understand a subset of $\mathcal{P}$, $\mathcal{D} \subset \mathcal{P}$. All members of $\mathcal{D}$ are supposed to be discrete random variables, all following the same distribution.

2.1 *The Probable Cause Case*

Assume that a suspect of the crime is obtained through non-DNA methods and that his DNA profile, $\mathcal{E}_{\mathrm{S}}$, matches the one of the true perpetrator. How much evidence should be assessed to such a match? That is, to what extent does the fact that the suspect has a profile matching $\mathcal{E}_{\mathrm{TP}}$ affect the belief that this person is actually the true perpetrator? There is a broad consensus in the forensic community (Berry, 1991; Roeder, 1994) that the answer to that question is the value of a likelihood ratio related to two competing hypotheses, each describing the data under the following two scenarios:

(1) the prosecution's scenario, i.e., the suspect and the true perpetrator are one and the same person,
(2) the defense's scenario, i.e., the suspect and the true perpetrator are different persons.

These scenarios have the disadvantage that they do not describe the data (i.e., $\mathcal{E} = (\mathcal{E}_{\mathrm{TP}}, \mathcal{E}_{\mathrm{S}})$) in sufficient detail to allow us to assign a probability to the event $\{\mathcal{E} = (A, A)\}$ (i.e., matching profiles with a certain set of locus types (DNA profile) A). When we specify that the suspect and the true perpetrator are different persons, we do not specify in which way they are different. The usual alternative that is applied is that the suspect and the true perpetrator are independent or, more precisely stated, that their profiles $\mathcal{E}_{\mathrm{TP}}$ and $\mathcal{E}_{\mathrm{S}}$ are obtained independently from the population of profiles, considering the true perpetrator as a random man, sampled at random from the population (of persons) in question. Their marginal distributions are supposed to be identical, and taking the (theoretical) frequency of A as $p = p_A$, we can calculate the likelihood ratio between the two relevant hypotheses,

$\mathrm{H_p}$: $\mathcal{E}_{\mathrm{TP}}$ and $\mathcal{E}_{\mathrm{S}}$ are the same random variable

$\mathrm{H_d}$: $\mathcal{E}_{\mathrm{TP}}$ and $\mathcal{E}_{\mathrm{S}}$ are stochastically independent,

as

$$LR = \frac{P_{\mathrm{p}}(\mathcal{E}_{\mathrm{TP}} = A, \mathcal{E}_{\mathrm{S}} = A)}{P_{\mathrm{d}}(\mathcal{E}_{\mathrm{TP}} = A, \mathcal{E}_{\mathrm{S}} = A)} = \frac{P_{\mathrm{p}}(\mathcal{E}_{\mathrm{TP}} = A)}{P_{\mathrm{d}}(\mathcal{E}_{\mathrm{TP}} = A) P_{\mathrm{d}}(\mathcal{E}_{\mathrm{S}} = A)} = \frac{p}{p^2} = \frac{1}{p}, \qquad (1)$$

where we use subscripts p and d to denote the hypotheses put forth at a trial by the prosecution and the defense, respectively.

Remark. In practice, a modification of the product rule for independence is often used when the likelihood ratio (1) has to be calculated to account for relatedness within $\mathcal{P}$. This will not be discussed here (cf., Balding and Nichols, 1994; Weir, 1996a).

The hypotheses $\mathrm{H_p}$ and $\mathrm{H_d}$ may be given a meaning prior to the actual typing of the trace left at the crime scene and the suspect, i.e., the observation of $\mathcal{E}_{\mathrm{TP}}$ and $\mathcal{E}_{\mathrm{S}}$. If, in formula (1), the event $\{\mathcal{E}_{\mathrm{S}} = A\}$ is substituted by $\{\mathcal{E}_{\mathrm{S}} = B\}$ and A and B vary among the possible profiles, one obtains the likelihood ratios $1/p$ if $\mathcal{E}_{\mathrm{TP}} = \mathcal{E}_{\mathrm{S}}$ and zero if $\mathcal{E}_{\mathrm{TP}} \neq \mathcal{E}_{\mathrm{S}}$.

2.2 *The Database Search Case*

Consider the case where no suspect has been identified on the basis of non-DNA evidence and a database $\mathcal{D}$ of size n is subsequently searched in order to link a person to $\mathcal{E}_{\mathrm{TP}}$. A lone match is obtained, i.e., exactly one profile in $\mathcal{D}$ is found to match $\mathcal{E}_{\mathrm{TP}}$. The person identified by the matching profiles becomes a suspect. As in the case with a suspect identified through non-DNA evidence, we would like to assign a weight to this profile match evidence. Suppose that the suspect happens to be Smith. The hypotheses

$\mathrm{H'_p}$: Smith and the true perpetrator are one and the same person
$\mathrm{H'_d}$: Smith and the true perpetrator are different persons

are in this case data-dependent statements since we had no way of knowing prior to the search that Smith would be the person that matched. If the hypotheses $\mathrm{H'_d}$ and $\mathrm{H'_p}$ were to describe the experiment exactly, it would correspond that Smith is under suspicion of having committed the crime, so he and the true perpetrator (i.e., the trace from the crime scene) are typed and then the rest of the database is searched for other matches to $\mathcal{E}_{\mathrm{TP}}$.

This way, the hypotheses claim to know the suspect *before* the search, and the likelihood ratio would then correspond to a single two-person comparison. The likelihood ratio associated with $\mathrm{H'_p}$ and $\mathrm{H'_d}$ would attain a value of zero if it turned out that the profile of Jones matched $\mathcal{E}_{\mathrm{TP}}$ and Smith's profile did not. This is, of course, an incorrect evaluation of the evidence, so if we want to describe 'conducting a database search' experiment under the two scenarios, we must, in general, refer to 'a suspect identified from the database' instead of 'Smith.' What is meant by a 'suspect identified from the database' is a person who is not exonerated by the database search, i.e., a person whose DNA profile is in $\mathcal{D}$ and matches the one of the TP, and the case outlined at the start of this subsection is

the case where a search of $\mathcal{D}$ results in a single suspect. The authorities searching $\mathcal{D}$ cannot neglect any match with $\mathcal{E}_{\mathrm{TP}}$ since it puts the person in question under suspicion, and the DNA evidence against him/her has to be evaluated.

The technical hypotheses that have to be set up are therefore hypotheses that describe the data subject to the scenarios that one of the suspects is identical to the true perpetrator versus that the profiles in $\mathcal{D}$ and $\mathcal{E}_{\mathrm{TP}}$ are the result of i.i.d. random variables. We do not have any way of knowing that there's only one suspect (i.e., only one profile in $\mathcal{D}$ matching $\mathcal{E}_{\mathrm{TP}}$), so the hypotheses are

$\mathrm{H_p}$: the true perpetrator is among the suspects identified from the database

$\mathrm{H_d}$: $\mathcal{E}_{\mathrm{TP}}$ is stochastically independent of the profiles in the database.

Note that the first hypothesis states that the profile $\mathcal{E}_{\mathrm{TP}}$ of the true perpetrator is a member of $\mathcal{D}$ and that the other hypothesis states that the true perpetrator is not in $\mathcal{D}$. An underlying assumption that is suppressed in the formulation of the hypotheses is that the profiles in $\mathcal{D}$ not identical to $\mathcal{E}_{\mathrm{TP}}$ are stochastically independent of each other and of $\mathcal{E}_{\mathrm{TP}}$. This means that we do not deal with the case where $\mathcal{D}$ contains (profiles of) related persons, e.g., brothers, nor with the case where $\mathcal{D}$ contains (profiles of) persons related to the true perpetrator. These situations will have to be dealt with separately.

Suppose that a search has been conducted and k profiles matching $\mathcal{E}_{\mathrm{TP}}$ are found. Denote the number of suspects found in $\mathcal{D}$ by $r_{\mathcal{D}}$, so that k is the realisation of $r_{\mathcal{D}}$. The likelihood ratio of $\mathrm{H_p}$ against $\mathrm{H_d}$ becomes

$$\begin{aligned} LR &= \frac{P_{\mathrm{p}}(\mathcal{E}_{\mathrm{TP}} = A, r_{\mathcal{D}} = k)}{P_{\mathrm{d}}(\mathcal{E}_{\mathrm{TP}} = A, r_{\mathcal{D}} = k)} \\ &= \frac{P_{\mathrm{p}}(r_{\mathcal{D}\setminus\{\mathcal{E}_{\mathrm{TP}}\}} = k-1 \mid \mathcal{E}_{\mathrm{TP}} = A) P_{\mathrm{p}}(\mathcal{E}_{\mathrm{TP}} = A)}{P_{\mathrm{d}}(r_{\mathcal{D}} = k \mid \mathcal{E}_{\mathrm{TP}} = A) P_{\mathrm{d}}(\mathcal{E}_{\mathrm{TP}} = A)} \\ &= \frac{\binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} \cdot p}{\binom{n}{k} p^{k}(1-p)^{n-k} \cdot p} = \frac{k}{np}. \end{aligned} \tag{2}$$

When only a single match is observed, formula (2) reduces to the one over n rule, $LR = 1/np$. In this case, The factor $1/n$ compared to the probable cause likelihood ratio reflects that the experiment carries out n two-person comparisons, with the same number of matches (one) as in the probable cause setting. Every member of the database is compared with $\mathcal{E}_{\mathrm{TP}}$, and thus all the corresponding persons are considered as suspects (in the usual sense) of the crime in question prior to the search of the database. But the corresponding probable cause likelihood ratios are zero for all but the one person matching.

Collins and Morton (1994) and Morton (1997) claim that the relevant event one should assign a probability is that at least one match is observed. This is not the point of view taken here, but apart from this difference, the authors obtain formulas for the weight of the evidence that are approximately the same as (2).

A few more things should be noted about formula (2).

(1) It is the likelihood ratio of two hypotheses that may be formulated prior to the experiment, describing the process of considering persons with matching profiles in the database as suspects, which is what actually happens.

(2) When only the likelihood theory is considered and only the experiment 'conducting a database search' is described, the different approaches to the problem are limited to variations on the dependence/independence between different members of $\mathcal{D}$ and the true perpetrator and the incorporation of non-DNA evidence; formula (2) is the unique likelihood ratio derived from the likelihood theory.

(3) It takes proper account for the $n-1$ persons being excluded since this is a part of the observations evaluated.

(4) It is in concordance with the recommendations in the NRC II report, and it is derived solely from the likelihood theory.

3. Taking the Sampling Process into Account; Finite Populations

Suppose that the size of $\mathcal{P}$ is N and is thus not infinite and that $\{\mathcal{E}_{\mathrm{TP}} = A\}$ and exactly one match in $\mathcal{D}$ is observed. This situation is what is sometimes referred to as 'the island problem' (Dawid, 1994; Balding and Donnelly, 1995b; Dawid and Mortera, 1996) since it corresponds to the case where the crime has been committed on an isolated island, which the true perpetrator has had no opportunity to leave. Thus, that $\mathcal{P}$ is finite means that, compared to the infinite population case, we have the additional information that the true perpetrator belongs to a limited finite population. It is easy to check that this information changes neither the arguments nor the results from the section above as long as the likelihood ratio is calculated with respect to the same hypotheses.

An allegory that is often used to justify the use of $1/p$ instead of $1/np$ is that there is a difference in finding exactly one match when n corresponds to a small part of the population and finding exactly one match after having searched all of the population but one person, corresponding to $n = N - 1$, still assuming that the population in question does contain the true perpetrator. Though the number of persons checked is the same in the two cases, the $1/n$ rule is not reasonable to use in the last case, and it would therefore be meaningless to divide by the database size in order to correct for the chance of DNA profiles matching $\mathcal{E}_{\mathrm{TP}}$ by chance.

This is not a paradox. The point is that when we claim that the $1/n$ rule is not reasonable to use in the last case it is because we believe that there is a certain chance that we have sampled the true perpetrator when we created our database of size $N-1$ and that this chance is big. This means that we choose to describe the sampling process as well as the actual matches so that we first sample $\mathcal{D}$ at random from $\mathcal{P}$ and afterwards observe $\mathcal{E}_{\mathrm{TP}}$ and the number of matches in $\mathcal{D}$. Compared to the cases treated earlier, it means that we are formally adopting a Bayesian framework since we insist on assigning a probability to the hypotheses $\mathrm{H_p}$ and $\mathrm{H_d}$. The proper way to evaluate the evidence is therefore to assign *a priori* probabilities to $\mathrm{H_p}$ and $\mathrm{H_d}$ and to compute the posterior odds of the hypotheses given the data.

For example, one may have a mass screening situation, where a large part of a population has been investigated, which makes it unlikely that the true perpetrator by chance isn't in the database.

Using the notation p, n, and $r_{\mathcal{D}}$ from the earlier section, the chance of sampling $\mathcal{E}_{\mathrm{TP}}$ at random in $\mathcal{D}$ from $\mathcal{P}$, using the uniform distribution on the subsets of $\mathcal{P}$ of size $|\mathcal{D}|$, is n/N, while the probability of not sampling $\mathcal{E}_{\mathrm{TP}}$ is $1 - n/N$. The posterior odds for the hypotheses $\mathrm{H_p}$ and $\mathrm{H_d}$ is therefore, after observing the events $\{r_{\mathcal{D}} = k\}$ and $\{\mathcal{E}_{\mathrm{TP}} = A\}$,

$$\begin{aligned}\text{posterior odds} &= \frac{P(\mathcal{E}_{\mathrm{TP}} \in \mathcal{D})}{\mathcal{P}(\mathcal{E}_{\mathrm{TP}} \notin \mathcal{D})} \frac{\binom{n-1}{k-1} p^k (1-p)^{n-k}}{\binom{n}{k} p^{k+1} (1-p)^{n-k}} \\ &= \frac{\frac{n}{N}}{1 - \frac{n}{N}} \frac{k}{np} \\ &= \frac{k}{(N-n)p}. \end{aligned} \tag{3}$$

Thus, in the case of a single match, the posterior odds differs from the probable cause likelihood ratio by a factor $1/(N-n)$, i.e., one over the size of the database, where the true perpetrator's profile is present under the hypothesis $\mathrm{H_d}$. This factor reflects the odds of sampling the true perpetrator at random, $n/(N-n)$, a factor that increases with the size of the database, divided by the number of comparisons made, n.

Formula (3) was noted by Balding and Donnelly (1995b), but instead of using it as the weight of the evidence, Balding and Donnelly use it to argue that the evidence is not "n times less convincing than had the suspect been the only individual examined." To assign the weight of $1/np$ to the evidence, as was done in the previous section, is therefore claimed to be "misleading" (Balding and Donnelly, 1995b, p. 30) because the posterior odds of $\mathrm{H_p}$ versus $\mathrm{H_d}$ increases with n. Furthermore it is noted that the strength of the evidence depends only weakly on n when $n \ll N$. These points are certainly true when each member of $\mathcal{P}$ is assigned the same prior probability of belonging to the database $\mathcal{D}$, but Balding and Donnelly do not carry their argument through to assigning the much lower weight $1/(N-1)p$ to the evidence in the situation where the suspect is the only individual examined, which is correct if only one person is searched and a uniformly distributed sampling process is taken into account. Their formulations suggest that the case where the suspect is the only individual examined should be evaluated via the probable cause likelihood ratio $1/p$ in the case of a match. In doing this, they fail to distinguish between the probable cause case, where a uniform prior is inappropriate, and the case of searching a database of size one.

Remark. It is important to stress that the prior belief that the true perpetrator is in the database is based solely on the size of the database. This may apply when a database is constructed after a crime has been committed, based on a neutral selection criteria, which could be, e.g., place of residence. When the database is constructed before the crime is committed, it is not so obvious how the prior belief is to be assessed. One could argue that the fact itself that the people are in the database increases the prior belief that the true perpetrator is among them, e.g., if the database consists of convicted criminals. This is a dangerous route to take since it involves factors like social behaviour and the psychology of criminals, and under any circumstances, one should keep in mind that it means that one has an increased prior belief that the persons in question have committed a particular crime that is in question. This is not the same as having a prior belief that a convicted criminal may commit (or has committed) other crimes since it isn't clear that committing a crime under one set of circumstances should necessarily imply an increased chance of committing a crime under another set. Circumstances like the type of the crime or geography would play a role here.

Formula (3) may be used to measure the consequences of not taking the database sizes into account. Formula (3) becomes $1/p$ if $k = 1$ and $n = N - 1$, meaning that all of the population but one is searched and a single match is the result. When database sizes are not taken into account, the weight that the DNA evidence is given therefore corresponds to the situation where all of the population of possible perpetrators except for one is searched and the suspect is the only match. Needless to say, a situation like this will be very incriminating for any suspect as long as the database size is just moderate. The analysis above shows that it is possible to quantify in a reasonable way situations where $\mathcal{D}$ consists of a large part of $\mathcal{P}$, and instead of ignoring the database size in this case, one should work with a sliding correction factor to $1/p$ that ranges from $1/N$ to 1, and only reaches one in the extreme case where all of $\mathcal{P}$ but one is searched. It is not surprising that the coefficient is one only in this case since this is the only case that corresponds to a single two-person comparison; the person that has not been typed is the random man who is used to compare with in the probable cause scenario, where also only a single comparison is made.

4. Discussion and an Example

Several other papers have addressed the question of assessing the strength of the evidence of a unique match in a database search. In Balding and Donnelly (1996), the authors define the DNA evidence to be $\mathcal{E}_{\mathrm{TP}}$, $\mathcal{E}_{\mathrm{S}}$, and the event that none of the other individuals (than the suspect) matches $\mathcal{E}_{\mathrm{TP}}$ (the event O in their notation). They calculate a likelihood ratio of the DNA evidence subject to the condition (hypothesis) that the suspect is identical to the true perpetrator versus that the suspect is not the true perpetrator. In their calculations, they consider the suspect as a fixed person (i.e., Smith) and thus take the route outlined in Section 2.2 as the hypotheses H'_{p} and H'_{d}. They end up with the probable cause likelihood ratio, with a correction factor that is claimed to be slightly greater than one (the factor is the ratio of the probability of the event O subject to the two hypotheses when the suspect is considered as a fixed person). When the authors claim that this likelihood ratio is the appropriate strength of the DNA evidence, they grossly overstate the strength of the evidence, as described in Section 2.2. As noted there, their problem is that they formulate their hypotheses after the experiment has been conducted and let the hypotheses depend on the outcome. If, e.g., Brown had turned out to match instead of Smith, their approach would require that the hypothesis 'Brown is the true perpetrator' be formulated and weighted against the alternative 'Brown is not the true perpetrator.' Balding and Donnelly (1996) acknowledge that arguments corresponding to the hypotheses $\mathrm{H_p}$ and $\mathrm{H_d}$ in

Section 2.2 lead to a reduction of the strength of the evidence by a factor corresponding to the database size, but they claim that this quantity is irrelevant since 'at trial a court is concerned only with the suspect, and not with the collective guilt or innocence of the database.' What they refer to as 'collective guilt or innocence' is better named as 'collectively under suspicion,' which is exactly what is investigated by the authorities when each profile in a database is compared with the profile of the true perpetrator, and the court is, of course, concerned with an increased possibility of a random match. The decision problem of the court should take the implications of statistical hypotheses for data description into account and not the other way around, which is what Balding and Donnelly (1996) do.

An example that is being discussed goes like this: Suppose that a suspect is found through non-DNA evidence, which means that the appropriate evaluation of the DNA evidence is the probable cause likelihood ratio. We refer to this person as the primary suspect. Afterwards, a database search is conducted to see if any other profiles in a relevant database matches, and no other match is found. The reason for searching the database is to see if any of the members matches $\mathcal{E}_{\mathrm{TP}}$ and, consequently, to put any person with a matching profile in the database under suspicion. The relevant hypotheses in this case are therefore a combination of the two in Section 2.1 and 2.2, namely,

$\mathrm{H_p}$: The true perpetrator is either the primary suspect or one of the suspects in the database

versus

$\mathrm{H_d}$: The true perpetrator's profile is stochastically independent of both the profile of the primary suspect and of the persons in the database.

Again, these are two competing hypotheses, with $\mathrm{H_p}$ composite and $\mathrm{H_d}$ simple, and since the two experiments are carried out independently of each other, the likelihood ratio factors into two factors, where the first is the probable cause likelihood ratio and the second is the likelihood ratio corresponding to the database search. Since neither the true perpetrator nor the suspect is in the database if no match is found during the search, the maximised probability of no match in the database is the same subject to the two hypotheses, and the corresponding likelihood ratio is one. Again, the result is equal to the probable cause likelihood ratio.

The situation is confused by suggestions in Balding and Donnelly (1996) that 'a cunning defense lawyer' could claim that a database search has been conducted and the case against his client has been 'substantially weakened' (i.e., one should divide with the database size). On the other hand, if a second match was found during the database search, the case against his client would be weakened as well. No matter the outcome of the search, the defense has a better case afterward. This so-called 'cunning defense lawyer' fails to take the independence relation between the two experiments into account, and this is another example of the importance of realising what the actual circumstances of the database search are.

The assessment of *a priori* probabilities based on a uniform sampling procedure in the form considered in Section 3 is not relevant here because of the inclusion of the primary suspect, so other strategies have to be considered if one insists on using such probabilities. The use of general prior probabilities of the suspect actually being the true perpetrator is not recommended here on the grounds noted in the remark in Section 3, but it is important to realise that the approach does not support our 'cunning defense lawyer' in any way.

Suppose the primary suspect has a prior probability r of being the true perpetrator and assign uniform *a priori* probabilities $(1-r)/(N-1)$ to the rest of the population in question, where as in Section 3, N is the size of $\mathcal{P}$, so that we first assign the TP to one of the N possible perpetrators according to this distribution and afterward pick a database $\mathcal{D}$ of size n from the $N-1$ persons not being the primary suspect according to a uniform selection procedure. Then the two experiments are performed: First, the profiles of the TP and the primary suspect are compared. Second, the database is searched. If the results are matching profiles from the TP and the primary suspect and no matches to their common profile in the database, the posterior odds of the primary suspect being the TP, after the first experiment is conducted, is

$$\text{posterior odds} = \frac{1}{p} \cdot \frac{r}{1-r}, \tag{4}$$

while the posterior odds of the hypotheses $\mathrm{H_p}$ and $\mathrm{H_d}$ formulated above after conducting the second experiment is

$$\begin{aligned}
&\text{posterior odds}\\
&= \frac{P(\mathcal{E}_{\mathrm{TP}} = \mathcal{E}_{PS}, r_{\mathcal{D}} = 0 \mid TP = PS)P(TP = PS) + P(\mathcal{E}_{\mathrm{TP}} = \mathcal{E}_{PS}, r_{\mathcal{D}} = 0 \mid \mathcal{E}_{\mathrm{TP}} \in \mathcal{D})}{P(\mathcal{E}_{\mathrm{TP}} = \mathcal{E}_{PS}, r_{\mathcal{D}} = 0 \mid TP \neq PS \text{ and } \mathcal{E}_{\mathrm{TP}} \notin \mathcal{D})P(TP \neq PS \text{ and } \mathcal{E}_{\mathrm{TP}} \notin \mathcal{D})}\\
&= \frac{1}{p} \cdot \frac{P(TP = PS)}{P(TP \neq PS)} \cdot \frac{1}{P(\mathcal{E}_{\mathrm{TP}} \notin \mathcal{D} \mid \mathcal{TP} \neq \mathcal{PS})}\\
&= \frac{1}{p} \cdot \frac{r}{1-r} \cdot \frac{N-1}{N-n-1},
\end{aligned} \tag{5}$$

where the PS and $\mathcal{E}_{PS}$ is shorthand for the primary suspect and the corresponding DNA profile, respectively. Now, (5) is $(N-1)/(N-n-1)$ times (4), meaning that the database search in this case increases the strength of the evidence to $(N-1)/(N-n-1)$ times the strength in the probable cause scenario, irrespective of the prior probability r. The posterior odds is therefore higher than in the probable cause case, as it should be, but as long as the database $\mathcal{D}$ does not consist of nearly all of $\mathcal{P}$, the factor is close to one.

5. Conclusion

The impact on the likelihood ratio of the size of the database searched has been demonstrated with the difference of the formulas for the probable cause likelihood ratio (1) and the standard database search likelihood ratio (2). If a database is searched and the population of possible perpetrators is a limited, finite population, the posterior odds (3) that takes the sampling process into account can be assigned as the weight of the evidence, provided that a uniform prior applies. This weight will be larger than the weight obtained as the standard database search likelihood ratio (2) if and only if more than half of the population is searched. The posterior odds (3) is smaller than the probable cause likelihood ratio (1) with a factor equal to the size of the part of the population that is not searched and thus is equal to (1) if and only if all but one of the population is searched. The standard database search likelihood ratio (2) is smaller than the probable cause likelihood ratio (1), with a factor equal to the size of the database searched, with a corresponding decrease in the weight of the evidence as a consequence.

When a suspect of a crime is found as the result of a database search, the weight of the DNA evidence may be severely overstated if the database search is not taken into account. However, using the database search likelihood ratio may lead to the assessment of a weight to the evidence that makes it useless to the court. How should this be accounted for? The obvious answer is simply to type more loci. Assuming no linkage and no gametic association between the loci in question, types on new loci should be stochastically independent of those used to identify the suspect in the database, and the likelihood ratio for the new loci can be multiplied to the database search likelihood ratio. Alternatively, one could, before the database search is conducted, type additional loci in order to make the database search likelihood ratio large, but this procedure is limited by the number of typed loci stored in the database. The resulting likelihood ratio will be the same, and any suspect that it is possible to actually get hold of may be typed for additional loci afterwards.

The new information will either acquit the suspect or strengthen the evidence to a level where the correction factor from the database search does not lead to 'reasonable doubt.' With the number of loci that are available for forensic identification at present, this is not a problem if sufficient DNA from the trace from the crime scene is available. If this is not the case, the strength of the evidence on the basis of the database size in question should be evaluated, and if it is considered weak, it should be acknowledged that weak evidence is nothing but weak evidence. The database used by the Forensic Science Service in England consists presently of more than 250,000 persons, and it has been reported that the most common profiles appear multiple times. This definitely means that random database matches cannot be ignored. For a database of this size and a unique match in the database to a profile with a frequency of one in a million, the corresponding likelihood ratio would be less than four. And this certainly does reflect that, assuming that the true perpetrator is not in the database, in more than one out of four cases, there would be an innocent person in the database matching the profile with a frequency of one in a million by chance alone.

Acknowledgements

I would like to thank Dr Jørgen Hilden and Dr Joseph L. Gastwirth for stimulating discussions and Professors Niels Morling and Niels Keiding for useful suggestions on an earlier draft of the manuscript.

Résumé

Un crime a été commis, et l'empreinte DNA du coupable est obtenue sur les lieux du crime. Un suspect avec une empreinte correspondante est trouvé. Le problème de l'évaluation de cette évidence au sens du DNA dans un contexte légal, lorsque le suspect est trouvé par une recherche dans une base de données, est analysé par une approche de vraisemblance. Les recommandations du Conseil National de la Recherche des Etats-Unis sont obtenues dans ce contexte, comme la méthode correcte pour évaluer cette évidence lorsqu'on ne prend pas en compte la taille finie de l'ensemble des coupables possibles. Si l'on peut supposer cette population de taille finie, il est possible d'en tenir compte dans le processus d'échantillonnage dans la base de données afin qu'on puisse traiter le problème avec une forte proportion de coupables possibles appartenant à la base de données utilisée. On montre que cette dernière approche ne permet pas en général d'affecter un plus grand poids à l'évidence, bien qu'il en soit ainsi lorsqu'une quantité suffisante de coupables possibles soit dans la base de données. La valeur du rapport de vraisemblance correspondant au contexte de la cause probable est une borne supérieure de ce poids, et la borne supérieure est atteinte seulement lorsque tous les coupables possibles sauf un sont dans la base de données.

References

Aitken, C. G. G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester, U.K.: Wiley.

Balding, D. J. (1996). DNA profiles and the courts. *The Royal Statistical Society News* **24**, 1–2.

Balding, D. J. (1997). Errors and misunderstandings in the second NRC report. *Jurimetrics* **37**, 603–607.

Balding, D. J. and Donnelly, P. (1995a). Inferring identity from DNA profile evidence. *Proceedings of the National Academy of Science, USA* **92**, 11741–11745.

Balding, D. J. and Donnelly, P. (1995b). Inference in forensic identification. *Journal of the Royal Statistical Society, Series A* **158**, 21–53.

Balding, D. J. and Donnelly, P. (1996). Evaluating DNA profile evidence when the suspect is identified through a database search. *Journal of Forensic Sciences* **41**, 603–607.

Balding, D. J. and Nichols, R. A. (1994). DNA profile match probability calculation: How to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International* **64**, 125–140.

Balding, D. J., Donnelly, P., and Nichols, R. A. (1994). Some causes for concern about DNA profiles. Comment in Roeder, K. (1994). DNA fingerprinting: A review of the controversy. *Statistical Science* **9**, 248–251.

Berry, D. (1991). Inferences using DNA profiling in forensic identification and paternity cases. *Statistical Science* **2**, 175–205.

Berry, D. (1994). Comment in Roeder, K. (1994). DNA fingerprinting: A review of the controversy. *Statistical Science* **9,** 252–255.

Collins, A. and Morton, N. E. (1994). Likelihood ratios for DNA identification. *Proceedings of the National Academy of Science* **91,** 6007–6011.

Dawid, A. P. (1994). The island problem: Coherent use of identification evidence. In *Aspects of Uncertainty: A Tribute to D. V. Lindley,* P. R. Freeman and A. F. N. Smith (eds), Chapter 11. Chichester, U.K.: Wiley.

Dawid, A. P. and Mortera, J. (1996). Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society, Series B* **58,** 425–443.

Evett, I. W. and Weir, B. S. (1998). *Interpreting DNA Evidence.* Sunderland, Massachusetts: Sinaur.

Lempert, R. (1997). After the DNA wars: Skirmishing with NRC II. *Jurimetrics* **37,** 439–468.

Morton, N. E. (1997). The forensic DNA endgame. *Jurimetrics* **37,** 477–494.

National Research Council Committee on DNA Forensic Science, USA. (1992). *DNA Technology in Forensic Science.* Washington, D.C.: National Academy Press.

National Research Council Committee on DNA Forensic Science. (1996). *An Update: The Evaluation of Forensic DNA Evidence.* Washington, D.C.: National Academy Press.

Robertson, B. and Vigneaux, G. A. (1995). *Interpreting Evidence.* Chichester, U.K.: Wiley.

Roeder, K. (1994). DNA fingerprinting: A review of the controversy. *Statistical Science* **9,** 222–278.

Taroni, F. and Aitken, C. (1997). Forensic science at trial. *Jurimetrics* **37,** 327–337.

Weir, B. S. (1996a). *Genetic Data Analysis II.* Sunderland, Massachusetts: Sinaur.

Weir, B. S. (1996b). The second National Research Council report on forensic DNA evidence. *American Journal of Human Genetics* **59,** 497–500.

Received February 1998. *Revised July* 1998.
Accepted August 1998.