Dept. of Math. University of Oslo
Statistical Research Report No. 2
ISSN 0806–3842 May 2006

The DNA database search controversy revisited: Bridging the Bayesian – Frequentistic gap

Geir Storvik[*] Thore Egeland[†]

May 10, 2006

Abstract

Two different quantities have been suggested for quantification of evidence in cases where a suspect is found by a search through a database of DNA profiles. The likelihood ratio, typically motivated from a Bayesian setting, is preferred by most experts in the field. The so-called *np* rule has been suggested through more frequentistic arguments and has been suggested by i.e. American National Research Council and Stockmarr [1999]. The two quantities differ substantially and have lead to what is called the DNA database search controversy. Although several authors have criticized the different approaches, a full explanation of why these differences appear is still lacking.

In this paper we show that a quantity approximately equal to the *np* rule can be seen as a P-value in a frequentistic hypothesis setting. We argue however that a more reasonable procedure in this case is to use conditional testing, in which case a P-value directly related to posterior probabilities and the likelihood ratio is obtained. This way of viewing the problem bridge the gap between the Bayesian and frequentistic approaches. At the same time it indicates that the *np* rule should not be used as a quantity of evidence.

Keywords— **DNA, database search, conditioning, forensic genetics**

1 Introduction

A crime has been committed and a trace (i.e blood, semen saliva) is left on the crime scene, from which a DNA profile is recorded. No suspects are pointed out, but a database containing individuals with known DNA profiles is available. A search through this database is performed and *exactly one match* is found. The individual with the match is put to trial. How should one quantify the evidence?

In the case where a suspect has been pointed out by other means than the DNA profile, the accepted way to report the evidence is through the likelihood ratio between the hypothesis that the suspect is the source of the stain and its compliment. This ratio reduces to $\frac{1}{p}$ where p is the match probability of the given profile in the population of consideration [Balding and Donnelly, 1995, Dawid and Mortera, 1996, Evett and Weir, 1998, Balding, 2005].

In the DNA database search case, two different suggestions have been made for quantifying the evidence. On approach has been to report the likelihood ratio also in this case. Under

[*] Department of mathematics and Centre for Ecological and Evolutionary Synthesis, University of Oslo, and Norwegian Computing Center, email geirs@math.uio.no

[†] Department of medical genetics, Ullevaal University Hospital, email Thore.Egeland@medisin.uio.no

simplified assumptions the likelihood ratio becomes $(1+\frac{n-1}{N-n})\frac{1}{p}$, where n is the size of the database and N is the population size. The arguments for using the likelihood ratio have both been its direct interpretation and its measure as the change of odds between hypotheses in a Bayesian framework. The other approach has been to report $\frac{1}{np}$ as the measure of evidence. This approach has been named the np rule and has been suggested both by the American National Research Council [NRC] and by Stockmarr [1999], though by different arguments.

The two approaches differ substantially. While the likelihood ratio strengthens the evidence with increasing n (although only slightly when $N >> n$), the np rule decreases the evidence dramatically with increasing n. Both camps have heavily criticized each others view with no agreement being made. The discussion between these approaches now has somewhat died out, with most scientists in the field preferring the likelihood ratio approach. Still, a satisfactory explanation for the differences between these approaches and full understanding with respect to some of the aspects in the criticisms that have appeared is lacking.

In this paper we will have another look on the database controversy by using frequentistic tools such as type I errors and P-values. Looking at the probability for type I error, proper account can be taken both for the multiple testing problem and data-dependent hypotheses, aspects that were the main arguments for the criticism by [NRC] and Stockmarr [1999]. The probability for type I error is in close relation to the np rule. We will however argue that a more proper setting in this case is to use a conditional testing procedure, in which case the probability for type I error corresponds to the Bayesian posterior probability. These results both bridge the gap between the different views on the DNA database search controversy and at the same time explains that the differences is rather related to conditioning than to other methodological issues.

The outline of this paper is as follows: In section 2 we review the database controversy and the main criticisms from both camps. In section 3 we consider the problem from a frequentistic point of view and derive type I probabilities and P-values. In section 4 we consider the conditional frequentistic approach while a summary and further comments are given in section 5.

2 The controversy

In the following we will assume the match probability p and the population size N are known. We will further assume that the DNA profiles are read without errors and that the DNA profiles of the individuals in the database are independent. All these assumptions are questionable in real settings, but are made in order to simplify the discussion. In practice the extra uncertainty introduced by the violation of these assumptions needs to be taken properly into account, see Balding [2005].

In order to discuss the problem from a statistical viewpoint, consider the hypotheses

$$H_j : \text{Individual } j \text{ is contributor} \tag{2.1}$$

for $j = 1, ..., N$. The problem can be formulated as a quantification of evidence towards H_s compared to the complement (that one of the other $H_j's$ is true).

The likelihood ratio between hypotheses H_s and its complement H_s^c for the database case is [Balding and Donnelly, 1995, Dawid and Mortera, 1996]

$$LR = (1+\tfrac{n-1}{N-n})\tfrac{1}{p}. \tag{2.2}$$

For $n = 1$, LR reduces to $\frac{1}{p}$. When n increases, so does LR, though only slightly if N is large compared to n. Defining odds in favour of H_s compared to its complement, we have

$$\text{Posterior odds} = \text{Prior odds} \times LR.$$

From a Bayesian point of view this relation can be used as an argument that the evidence towards H_s slightly increases when the size of the database n gets larger. This can also be seen through the posterior probability of H_s which, assuming $\Pr(H_s) = \frac{1}{N}$ (a conservative prior in this setting), is given by

$$\Pr(H_s|y) = \frac{1}{1 + (N - n)p}. \tag{2.3}$$

(Dawid [2001] considered more general priors).

The National Research Council [NRC] argued that when searching through a database of size n, there are n possibilities for match. The evidence should therefore be weakened by a factor n, giving $\frac{1}{np}$ as the weight of evidence. Donnelly and Friedman [1999] criticized the NRC approach by arguing that they ask the wrong question and fail to recognize the full import of evidence of identification based on a database search. Balding [2002] criticized the value $1/np$ and argued that it is not reasonable that searching through a database should give weaker evidence. The np rule does not take into account that only one match *excludes* $n - 1$ suspects and does not fit with the situation where $n = N$ which gives certain identification! He further commented on the issue on multiple testing, and claimed that this is not a problem in this case because only *one* of the H_j's can be true.

Stockmarr [1999] also criticized the use of LR given in (2.2), but from another point of view. His main argument was that the hypothesis being considered, H_s, depends on data, in which case the use of the likelihood ratio (2.2) directly is problematic. He therefore suggested the use of an alternative hypothesis,

$$H_p : \text{Contributor is in the database,}$$

which do *not* depend on data but conditionally on the data, becomes equivalent to H_s. Using similar assumptions as those behind (2.2), he then obtained a likelihood ratio of $\frac{1}{np}$ for H_p against H_p^c. This he used for arguing that the evidence is *weaker* when searching a database.

The paper by Stockmarr resulted in several letters to the editor in Biometrics and also other papers who all criticized his approach. One of the points made was that H_p/H_p^c are not the hypotheses of interest [Evett et al., 2000]. Dawid [2001] discussed this further making the point that even though H_s and H_p are conditionally equivalent, they are not so before the search and using the likelihood ratio for H_p/H_p^c to evaluate H_s/H_s^c does not make sense.

3 A frequentistic approach

Consider the problem from a frequentistic setting. In order to explain this approach, we will consider this in a proper hypothesis testing with significance levels and P-values. We do not claim that these are quantities that should be used in court, but will merely use them as tools for explanation.

We want to test $H_j, j = 1, ..., N$ as defined in (2.1). Consider a test-strategy where we accept H_s and reject $H_j, j \notin s$ if a single match at s is found *and* the probability for making a wrong decision is smaller than a constant α. As for likelihood ratio calculations for the H_p hypothesis, it is not possible to make the necessary calculations without further "prior" assumptions about the hypotheses H_j. In particular, we will assume the probability for the contributor being outside the database is $(N-n)/N$ (this could be generalized to other priors as in Dawid [2001]). We further follow Balding and Donnelly [1995] and Dawid and Mortera [1996] in assuming independence of DNA profiles between individual in the database. Let $\mathcal{D}$ be the set of individuals in the database. Then

$$\begin{aligned}
&\Pr(\text{Reject a true } H_s)\\
&\quad=\sum_i \Pr(H_i)\Pr(\text{Reject } H_i|H_i)\\
&\quad=\sum_{i\notin\mathcal{D}} \Pr(H_i)\Pr(\text{one match for } s\in\mathcal{D}|H_i)\\
&\quad=np(1-p)^{n-1}\sum_{i\notin\mathcal{D}}\Pr(H_i).
\end{aligned}$$

where we have used that one match at s is impossible if $H_i, i\in\mathcal{D}, i\neq s$ is true. A further specification is not possible without making some "prior" assumptions on the hypotheses H_i. Assuming the probability for the true hypothesis to be inside the database is n/N (a conservative choice with respect to the suspect), we obtain

$$\Pr(\text{Reject a true } H_s)=\frac{N-n}{N}np(1-p)^{n-1}. \tag{3.4}$$

We accept H_s and reject $H_j, j\notin s$ if this expression is lower than α. This give us a significance level for the test (less or) equal to α. Using that P-values can be defined as the smallest significance level for which rejection occur Goodman [2005], the P-value for the test coincide with (3.4). For large N compared to n and p small, this expression becomes close to np, indicating that at least in this setting the evidence towards H_s *decrease* with n. Note though that for $n=N$, we obtain a P-value of zero as we should because we then are certain about who is the contributor. One could also consider this through the H_p hypothesis. The probability of a false conclusion then coincide with the expression given in (3.4).

Comparing now the Bayesian approach with the frequentistic one, we get (almost) the same differences as between the Bayesian and the NRC/Stockmarr approaches. Now however a fully frequentistic approach is taken. Why this difference? In the next section we will demonstrate that agreement between the two approaches can be obtained by looking at the frequentistic approach in a conditional setting. This will simultaneously explain the differences described above.

4 A conditional frequentistic approach

Assume again a case is put to trial *if one match* is found in the database. The derivations on the probability of a type I error in the previous section were based on

- cases put to trial (one match) and,
- cases not put to trial (zero or many matches).

A low probability for a wrong decision could therefore be caused by a low probability for actually putting the case to trial (i.e. cases where exactly one match is not obtained). A more natural approach is to control the type I error *within cases put to trial.* Consider therefore a test-strategy where we reject H_s if a single match is found at individual s and the probability for making a wrong decision *among those cases put to trial* is lower than α. Then, under the assumptions made in section 2,

$$\begin{aligned}\Pr(\text{one match}) =&\sum_{i\in\mathcal{D}}\Pr(H_i)\Pr(\text{one match}|H_i)+\sum_{i\notin\mathcal{D}}\Pr(H_i)\Pr(\text{one match}|H_i)\\ =&\tfrac{n}{N}(1-p)^{n-1}+\frac{N-n}{N}np(1-p)^{n-1}.\end{aligned}$$

Combined with (3.4), we obtain

$$\Pr(\text{Reject a true } H_s|\text{one match}) = \frac{(N-n)p}{[1+(N-n)p]}, \tag{4.5}$$

which (using similar arguments as in the previous section) is the P-value for this conditional test. Note that in this case, for large N, the size of the database n is negligible and in fact the evidence *increases* slightly with n. Note further that the P-value is equivalent to the Bayesian posterior probability of H_s^c being true. An agreement between the frequentistic and the Bayesian approach is therefore now achieved!

5 Summary and further comments

The main conclusion of this work was to demonstrate that the same value on quantity of evidence can be obtained both from a frequentistic (using P-values) and a Bayesian (using posterior probabilities) point of view.

All derivations are based on assuming that the match probability p and the population size N are known and that the DNA profiles for the individuals are independent. Although such assumptions are questionable, taking dependence into account would only complicate the discussion and not change the main conclusion.

The derivations of the frequentistic P-values are actually not purely frequentistic in that the calculations are based on assumptions about the "prior" probabilities for the hypotheses H_j. This is of no concern to us since our main goal was just to show that the Bayesian solution also can be seen as reasonable from a frequentistic point of view where the concerns made by NRC and Stockmarr [1999] are properly taken into account.

An interesting observation is that by looking at P-values, the unconditional approach actually gives a *smaller* P-value than in the conditional case ((3.4) is always smaller than (4.5)). This is in striking contrast to the conclusions both NRC and Stockmarr [1999] made from the np rule.

Acknowledgment

The first author was in this work sponsored by the Norwegian Research Council through the project "Evaluation of Bayesian Hierarchical Models", project number 154911.

References

D. J. Balding. The DNA database controversy. *Biometrics*, 58:241–244, 2002.

D. J. Balding. *Weight-of-evidence for Forensic DNA profiles.* Statistics in practice. Wiley, 2005.

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

A. P. Dawid. Comment on stockmarr's "likelihood ratios for evaluating DNA evidence when the suspect is found through a database search". *Biometrics*, 57:976–978, 2001.

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

P. Donnelly and R. D. Friedman. DNA database searches and the legal consumption of scientific evidence. *Michigan Law Review*, 97:931–984, 1999.

I. W. Evett, R. D. Foreman, and B. S. Weir. Letter to the editor. *Biometrics*, 56:1274–1275, 2000.

I. W. Evett and B. S. Weir. *Interpreting DNA Evidence.* Sunderland, Massacusetts: Sinauer, 1998.

S. N. Goodman. P value. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics.* Wiley, second edition, 2005.

National Research Council (NRC). The evaluation of forensic DNA evidence. Technical report, National Academy Press, Washington D. C., 1996.

A. Stockmarr. Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. *Biometrics*, 55:671–677, 1999.