

# DATABASE LIKELIHOOD RATIOS AND FAMILIAL DNA SEARCHING

KLAAS SLOOTEN AND RONALD MEESTER

ABSTRACT. Familial Searching is the process of searching in a DNA database for relatives of a given individual, called the target. It is well known that in order to evaluate the genetic evidence in favour of a certain given form of relatedness between two individuals, one needs to calculate the appropriate likelihood ratio, which is in this context called a Kinship Index. Suppose that the database contains, for a given type of relative, at most one related individual. Given prior probabilities for being the relative for all persons in the database, we derive the likelihood ratio for each database member in favour of being that relative. This likelihood ratio takes all the Kinship Indices between target and members of the database into account. We also compute the corresponding posterior probabilities. We then discuss two ways of selecting a subset from the database that contains the relative with a known probability, or at least a useful lower bound thereof. We discuss the relation between these approaches and illustrate them with Familial Searching carried out in the Dutch National DNA Database. Since this method applies to any situation where one wishes to retrieve a ‘special’ item out of a set of ‘generic’ items based on likelihood ratios, we have formulated the search strategies in this general context.

## 1. INTRODUCTION

Many countries maintain databases that contain forensic DNA profiles of traces and of certain known individuals, e.g. convicted offenders or suspects of certain crimes. These databases were originally set up to directly identify an unknown offender by looking for matching DNA profiles. However, since DNA is inherited from parent to child, it is also possible to use them to look for the offender’s relatives, rather than the offender himself, if the offender’s DNA profile turns out not to be in the database. This last process is called familial (DNA) searching, and is carried out in several jurisdictions (e.g. the UK, some US states and New Zealand). As a result there have been some high profile successes (see e.g. [5] for the Grim Sleeper case). The Netherlands are currently preparing a law that will allow familial searching in some cases.

Previous studies on familial DNA searching have mostly concentrated on empirical determination of the rank that a relative (of some target profile) in the database occupies, when the database is ordered according to decreasing likelihood ratio with the target, or according to decreasing number of shared alleles with the target. See e.g. [1] for simulations that also includes a geographical component (based on US states), or [2], [6] and [4]. In [3], false exclusion rates and false inclusion rates

---

*Date:* October 17, 2011.

*Key words and phrases.* Familial Searching, Kinship Analysis, DNA-Databases, Bayesian Inference, Weight of Evidence.

for various thresholds on the likelihood ratio and/or number of shared alleles are estimated by simulation. These rates are averages over different target profiles.

In this paper we discuss a mathematical model in which we can interpret the likelihood ratios between the target and the database. Since our method is not specific to DNA, we present it in the more general setting where we have a target item that we compare with database items among which, we suppose, is at most one ‘related’ item present. If that special item is present in the database, we wish to find it, and we assume that we can compute one-to-one likelihood ratios between the target and every database member. We will compute (given prior probabilities) posterior probabilities for relatedness between the target and a database member that take *all* the computed one-to-one likelihood ratios into account.

This last feature is in contrast with the above mentioned articles, which have in common that all likelihood ratios with the database members are viewed in isolation, i.e., without taking account of the other likelihood ratios that have been obtained with the database. The only attempt to take into account the fact that a database search has been done, is described in [7] where the SWGDAM Ad Hoc Committee on Partial Matches recommends that the kinship index between a database member and the target be divided by  $N$ , the size of the database, and to only further investigate this possible lead if that quotient is sufficiently large. As we shall see, this number does not represent the likelihood ratio in favour of relatedness.

A related topic, but focussing on a specific target profile, is discussed in [8], where the situation is addressed that a suspect is excluded as a crime stain donor, but shares unusually many alleles with it. It is pointed out in [8] that the distribution of the likelihood ratio between the crime stain donor and his (say) brother, depends on the profile of the crime stain donor and that this can be used to formulate a criterion whether or not to investigate a possible relationship between suspect and crime stain donor: the likelihood ratio has to be sufficiently large, in the sense that it arises with small enough probability as a result of the comparison of DNA profile of unrelated individuals. They work outside of the database context, but clearly this gives a method to decide who to further investigate based on database searches: everybody who has a likelihood ratio above the threshold. However, the probability with which a relative is detected in that set is not under control (as the authors also remark).

The set up of this paper as follows. In Section 2 we derive some general properties of likelihood ratios. These should be well known but we are not aware of an explicit reference. In Section 3 we apply these properties to compute likelihood ratios and posterior probabilities in favour of a database subset containing the special item, given all likelihood ratios with the target. In particular, we compute how many chance matches one expects: if the database does not contain the special item, then the expectation of the sum of all likelihood ratios between target and database is equal to the number of items in the database. In Section 4 we define and compare two search strategies: one that takes prior probabilities and observed likelihood ratios into account (we will call this the conditional method) and one that only takes observed likelihood ratios into account (we will call this the target-centered method). The conceptual differences between these two approaches are significant and we discuss them in detail. Finally, in Section 5 we apply these results to familial searching in the Dutch National DNA Database where we assess their power by

performing familial searches in this database with artificial target profiles and their simulated relatives.

## 2. GENERAL LIKELIHOOD RATIO PROPERTIES

Before we proceed, we first discuss certain general properties of likelihood ratios that we will use later on. The idea is that we have observed evidence (e.g. a DNA profile or a set of DNA profiles), that potentially can be explained by several hypotheses. We ultimately wish to assess which of these hypotheses has generated these data, or at least come to a probabilistic statement.

Let  $H$  be an hypothesis, and let the set of possible evidence be denoted by  $E$ , e.g. the set of DNA profiles, or of tuples of DNA profiles. The hypothesis  $H$  induces a probability distribution, denoted by  $\mathbf{H}$ , on  $E$  in a natural way: the probability of  $e$  is the probability of observing  $e$  if  $H$  is true. In what follows we identify a hypothesis with a random variable whose distribution is precisely this induced probability distribution on  $E$ . It turns out that the formulae become somewhat more transparent this way but we stress that this identification is just for convenience, not for any conceptual reason. Interpreted this way, it makes sense to write  $P(\mathbf{H} = e)$ , denoting the probability to observe  $e$  under the assumption that  $H$  is correct.

Similarly, the likelihood ratio with respect to hypotheses  $H_1$  and  $H_2$ , usually denoted by

$$\text{LR}(e) = \frac{P(e|H_1)}{P(e|H_2)}$$

reads, in the new notation, as

$$(2.1) \quad \text{LR}(e) = \text{LR}_{\mathbf{H}_1, \mathbf{H}_2}(e) = \frac{P(\mathbf{H}_1 = e)}{P(\mathbf{H}_2 = e)}.$$

When no confusion is possible, we omit the subscripts of  $\text{LR}$ . In what follows,  $\text{LR}(\mathbf{H})$  denotes the composition of  $\mathbf{H}$  and  $\text{LR}$ .

**Proposition 2.1.** *Suppose that for all  $e \in E$  we have*

$$(2.2) \quad P(\mathbf{H}_1 = e) > 0 \Rightarrow P(\mathbf{H}_2 = e) > 0.$$

*Then we have, for all  $x \geq 0$ ,*

$$\frac{P(\text{LR}(\mathbf{H}_1) = x)}{P(\text{LR}(\mathbf{H}_2) = x)} = x.$$

*Proof.* Denote the part of  $E$  on which the likelihood ratio takes value  $x$  by

$$E_x = \{e \in E \mid \text{LR}(e) = x\}.$$

We write

$$\begin{aligned} P(\text{LR}(\mathbf{H}_1) = x) &= \sum_{e \in E_x} P(\mathbf{H}_1 = e) \\ &= \sum_{e \in E_x} \text{LR}(e) P(\mathbf{H}_2 = e) \\ &= x \sum_{e \in E_x} P(\mathbf{H}_2 = e) \\ &= x P(\text{LR}(\mathbf{H}_2) = x). \end{aligned}$$

□

**Proposition 2.2.** *Under assumption (2.2) we have  $E(\text{LR}(\mathbf{H}_2)) = 1$ .*

*Proof.* We write, using (2.1),

$$\begin{aligned} E(\text{LR}(\mathbf{H}_2)) &= \sum_{e \in E} \text{LR}(e)P(\mathbf{H}_2 = e) \\ &= \sum_{e \in E} P(\mathbf{H}_1 = e) = 1. \end{aligned}$$

□

Proposition 2.2 can be interpreted as expressing that for every choice of likelihood ratio, there will always be chance matches (likelihood ratios in favour of  $\mathbf{H}_1$  whereas the data were generated by  $\mathbf{H}_2$ ), as long as (2.2) holds. Moreover, if we expect fewer chance matches, then these matches will be stronger to the effect that the expected likelihood ratio is constant.

### 3. DATABASE LIKELIHOOD RATIOS

In the model that we set up, the population consists of  $n$  individuals, and contains one “special” member among the remaining “generic” members. Let  $\mathbf{P}_1, \dots, \mathbf{P}_n$  be independent random variables that are all distributed as either  $\mathbf{H}_1$  or  $\mathbf{H}_2$ , such that exactly one of the  $\mathbf{P}_i$  (the “special” member) is distributed as  $\mathbf{H}_1$ . The index of the special member is random and denoted by  $\mathbf{R}$ . We assume that we know the “prior” probabilities  $P(\mathbf{R} = i)$  for all  $i$ . We also suppose that all  $\mathbf{P}_i$  are conditionally independent given  $\mathbf{R}$ , that is, given which of them is distributed as  $\mathbf{H}_1$ . This means that all population members are unrelated to each other and to the offender, except for  $\mathbf{R}$ .

In the DNA context,  $\mathbf{P}_1, \dots, \mathbf{P}_n$  represent the population’s DNA profiles, and the special member could be the offender or a (specific) relative that we are looking for, e.g., the offender’s father. In our set-up, the generic members are all unrelated to the offender.  $\mathbf{H}_2$  corresponds to the DNA profile of a generic member of the population, whereas  $\mathbf{H}_1$  is distributed as the DNA profile of the offender or a relative thereof.

The database is denoted by  $\mathcal{D}$  and has size  $N \leq n$ . For convenience, we assume that the individuals in  $\mathcal{D}$  have labels  $1, \dots, N$ . We write  $\mathbf{D} = (\mathbf{P}_1, \dots, \mathbf{P}_N)$  for the corresponding random vector, and

$$\text{LR}_{\mathbf{D}} = (\text{LR}_{\mathbf{H}_1, \mathbf{H}_2}(\mathbf{P}_1), \dots, \text{LR}_{\mathbf{H}_1, \mathbf{H}_2}(\mathbf{P}_N)),$$

for the random vector representing the likelihood ratios that we obtain from the database. Finally, we set

$$|\text{LR}_{\mathbf{D}}| = \sum_{i=1}^N \text{LR}_{\mathbf{H}_1, \mathbf{H}_2}(\mathbf{P}_i).$$

As before, we omit the subscripts of  $\text{LR}$ , since the likelihood ratio will always be with respect to  $\mathbf{H}_1$  and  $\mathbf{H}_2$ . Given realizations  $\mathbf{P}_i = e_i$ ,  $i = 1, \dots, N$ , we obtain a realization  $\text{LR}_{\mathbf{D}} = (\text{LR}(e_1), \dots, \text{LR}(e_N))$  and  $|\text{LR}_{\mathbf{D}}| = \sum_{i=1}^N \text{LR}(e_i)$ . We will use these likelihood ratios from the database to make probabilistic statements concerning the identity of  $\mathbf{R}$ . In the next chapter we will use these to define search strategies for  $\mathbf{R}$  in the database.

First we discuss how to arrive at likelihood ratios and probabilities conditional on the observed likelihood ratios  $\text{LR}_{\mathbf{D}} = (r_1, \dots, r_N)$ . Abbreviating this last vector by  $\mathbf{r}$ , we have the following result.

**Proposition 3.1.** For  $i = 1, \dots, N$ , we have

$$(3.1) \quad P(\mathbf{R} = i \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r}) = \frac{r_i P(\mathbf{R} = i)}{\sum_{k=1}^N r_k P(\mathbf{R} = k) + P(\mathbf{R} \notin \mathcal{D})}$$

and

$$(3.2) \quad P(\mathbf{R} = i \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r}, \mathbf{R} \in \mathcal{D}) = \frac{r_i P(\mathbf{R} = i)}{\sum_{k=1}^N r_k P(\mathbf{R} = k)}.$$

*Proof.* We first prove (3.1). The required probability is equal to

$$\frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i) P(\mathbf{R} = i)}{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r})},$$

which can be expanded as

$$\frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i) P(\mathbf{R} = i)}{\sum_{k=1}^N P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = k) P(\mathbf{R} = k) + P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \notin \mathcal{D}) P(\mathbf{R} \notin \mathcal{D})}.$$

Therefore, it is more attractive to consider the reciprocal, and we obtain

$$\begin{aligned} \frac{1}{P(\mathbf{R} = i \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r})} &= \sum_{k=1}^N \frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = k) P(\mathbf{R} = k)}{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i) P(\mathbf{R} = i)} \\ &+ \frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \notin \mathcal{D}) P(\mathbf{R} \notin \mathcal{D})}{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i) P(\mathbf{R} = i)}. \end{aligned}$$

Recall that all  $\mathbf{P}_i$  are conditionally independent given  $\mathbf{R}$ . This means that the last expression reduces to

$$\begin{aligned} &\sum_{k=1}^N \frac{P(\mathbf{LR}(\mathbf{P}_i) = r_i \mid \mathbf{R} = k) P(\mathbf{LR}(\mathbf{P}_k) = r_k \mid \mathbf{R} = k) P(\mathbf{R} = k)}{P(\mathbf{LR}(\mathbf{P}_i) = r_i \mid \mathbf{R} = i) P(\mathbf{LR}(\mathbf{P}_k) = r_k \mid \mathbf{R} = i) P(\mathbf{R} = i)} + \\ &+ \frac{P(\mathbf{LR}(\mathbf{P}_i) = r_i \mid \mathbf{R} \notin \mathcal{D}) P(\mathbf{R} \notin \mathcal{D})}{P(\mathbf{LR}(\mathbf{P}_i) = r_i \mid \mathbf{R} = i) P(\mathbf{R} = i)}. \end{aligned}$$

We claim that for  $k \neq i$ , we have

$$(3.3) \quad \frac{P(\mathbf{LR}(\mathbf{P}_i) = r_i \mid \mathbf{R} = k)}{P(\mathbf{LR}(\mathbf{P}_i) = r_i \mid \mathbf{R} = i)} = \frac{1}{r_i}$$

and

$$(3.4) \quad \frac{P(\mathbf{LR}(\mathbf{P}_k) = r_k \mid \mathbf{R} = k)}{P(\mathbf{LR}(\mathbf{P}_k) = r_k \mid \mathbf{R} = i)} = r_k.$$

To see this, note that given  $\mathbf{R} = k$ ,  $\mathbf{P}_k$  is distributed as  $\mathbf{H}_1$ , and  $\mathbf{P}_i$  is distributed as  $\mathbf{H}_2$ , and then use Proposition 2.1. For  $k = i$ , the corresponding term in the sum is equal to 1. We can also apply Proposition 2.1 to the last term since  $\mathbf{R} \notin \mathcal{D}$  implies that  $\mathbf{P}_i$  is distributed as  $\mathbf{H}_2$ . From all this it follows that

$$\frac{1}{P(\mathbf{R} = i \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r})} = \sum_{k=1}^N \frac{r_k P(\mathbf{R} = k)}{r_i P(\mathbf{R} = i)} + \frac{1}{r_i} \frac{P(\mathbf{R} \notin \mathcal{D})}{P(\mathbf{R} = i)},$$

and (3.1) follows.

The proof of (3.2) is similar, we only sketch the difference with the proof of (3.1). The probability in question is equal to

$$\frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i) P(\mathbf{R} = i)}{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r}, \mathbf{R} \in \mathcal{D})},$$

which can be expanded as

$$\frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i)P(\mathbf{R} = i)}{\sum_{k=1}^N P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = k)P(\mathbf{R} = k)}.$$

From this point on, the proof proceeds as above.  $\square$

**Remark 3.2.** Note that  $P(\mathbf{R} = i \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r})$  does not depend on the distribution of  $\mathbf{H}_1$  and  $\mathbf{H}_2$ .

It follows from Proposition 3.1 that, for any subset  $\mathcal{D}' \subset \mathcal{D}$

$$(3.5) \quad P(\mathbf{R} \in \mathcal{D}' \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r}) = \frac{\sum_{i \in \mathcal{D}'} r_i P(\mathbf{R} = i)}{\sum_{k=1}^N r_k P(\mathbf{R} = k) + P(\mathbf{R} \notin \mathcal{D})}.$$

In particular, with  $\mathcal{D}' = \mathcal{D}$  we obtain

$$\frac{P(\mathbf{R} \in \mathcal{D} \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r})}{P(\mathbf{R} \notin \mathcal{D} \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r})} = \frac{\sum_{i=1}^N r_i P(\mathbf{R} = i)}{P(\mathbf{R} \notin \mathcal{D})},$$

and the likelihood ratio in favour of  $\mathbf{R} \in \mathcal{D}$  is given by

$$(3.6) \quad \frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \in \mathcal{D})}{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \notin \mathcal{D})} = \frac{\sum_{i=1}^N r_i P(\mathbf{R} = i)}{P(\mathbf{R} \in \mathcal{D})}.$$

Note that the likelihood ratio depends on the prior probabilities.

**Corollary 3.3.** *In odds form, we obtain*

$$\frac{P(\mathbf{R} = i \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r})}{P(\mathbf{R} \neq i \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r})} = \frac{r_i P(\mathbf{R} = i)}{\sum_{k=1, k \neq i}^N r_k P(\mathbf{R} = k) + P(\mathbf{R} \notin \mathcal{D})},$$

and the likelihood ratio in favour of  $\mathbf{R} = i$  is given by

$$\frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i)}{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \neq i)} = \frac{r_i P(\mathbf{R} = i)}{\sum_{k=1, k \neq i}^N r_k P(\mathbf{R} = k) + P(\mathbf{R} \notin \mathcal{D})}.$$

In case where the alternative to  $\mathbf{R} = i$  is taken to be  $\mathbf{R} \notin \mathcal{D}$ , we have the following result.

**Corollary 3.4.** *We have*

$$\frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i)}{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \notin \mathcal{D})} = r_i.$$

*Proof.* The required likelihood ratio can be rewritten as

$$\frac{P(\mathbf{R} = i \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r})P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r})}{P(\mathbf{R} = i)} \frac{P(\mathbf{R} \notin \mathcal{D})}{P(\mathbf{R} \notin \mathcal{D} \mid \mathbf{LR}_{\mathcal{D}} = \mathbf{r})P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r})},$$

from which the term  $P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r})$  drops out. Substituting (3.3) and (3.5) with  $\mathcal{D}' = \mathcal{D}$  leads to the required result.  $\square$

In the case where the prior distribution of  $\mathbf{R}$  on  $\mathcal{D}$  is uniform, the above derived formulas simplify, and for convenience we include them here. Let  $\pi_{\mathcal{D}} = P(\mathbf{R} \in \mathcal{D})$  and  $P(\mathbf{R} = i) = \pi_{\mathcal{D}}/N$  for all  $1 \leq i \leq N$ .

We obtain, as special case of (3.6),

$$(3.7) \quad \frac{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \in \mathcal{D})}{P(\mathbf{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \notin \mathcal{D})} = \frac{1}{N} \sum_{i=1}^N r_i,$$

which is independent of the prior  $\pi_{\mathcal{D}}$ , contrary to the general case. In this uniform case, we see that the results  $\text{LR}_{\mathcal{D}} = \mathbf{r}$  favour  $\mathbf{R} \in \mathcal{D}$  if and only if the *average* likelihood ratio on  $\mathcal{D}$  is greater than one.

The posterior probability that  $\mathbf{R} = i$  is now equal to

$$P(\mathbf{R} = i \mid \text{LR}_{\mathcal{D}} = \mathbf{r}) = \frac{r_i}{\sum_{k=1}^N r_k + N \frac{1-\pi_{\mathcal{D}}}{\pi_{\mathcal{D}}}},$$

with corresponding likelihood ratio

$$(3.8) \quad \frac{P(\text{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i)}{P(\text{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \neq i)} = \frac{r_i}{\frac{\pi_{\mathcal{D}}}{N-\pi_{\mathcal{D}}} (\sum_{k=1, k \neq i}^N r_k) + \frac{N(1-\pi_{\mathcal{D}})}{N-\pi_{\mathcal{D}}}}.$$

In the even more specific case that  $\pi_{\mathcal{D}} = 1$ , i.e., the database surely contains a relative and any of the members can be the relative with equal a priori probability, we simply get

$$P(\mathbf{R} = i \mid \text{LR}_{\mathcal{D}} = \mathbf{r}) = \frac{r_i}{\sum_{k=1}^N r_k}$$

and

$$\frac{P(\text{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i)}{P(\text{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \neq i)} = \frac{r_i}{\frac{1}{N-1} \sum_{k \neq i} r_k}.$$

**Example 3.5.** We can view the process of searching for a match with a DNA profile as a special case. In that case  $\mathbf{R}$  corresponds to the trace donor, and  $\mathbf{H}_1$  can only take value  $e_0$ , the DNA profile in question. On the other hand  $\mathbf{H}_2$  can take more values with probabilities given by the profile population frequencies. Let  $p$  denote  $P(\mathbf{H}_2 = e_0)$ , the random match probability of the profile  $e_0$ . In this situation,  $\text{LR}(\mathbf{P}_i)$  can take value 0 or  $1/p$ . The variables  $\mathbf{P}_1, \dots, \mathbf{P}_N$  correspond to the members of a database in which we look for the profile  $e_0$ . Suppose that the  $i^{\text{th}}$  database member is the only one that matches. Then  $\text{LR}(\mathbf{P}_i) = 1/p$  and all other  $\text{LR}(\mathbf{P}_k) = 0$  (for  $1 \leq k \leq N, k \neq i$ ), so the likelihood ratio (3.8) in favour of  $\mathbf{R} = i$  becomes

$$\frac{P(\text{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} = i)}{P(\text{LR}_{\mathcal{D}} = \mathbf{r} \mid \mathbf{R} \neq i)} = \frac{1}{p} \frac{N - \pi_{\mathcal{D}}}{N(1 - \pi_{\mathcal{D}})}.$$

If  $\pi_{\mathcal{D}} = N/n$  (the population fraction in the database), then this reduces to  $(n-1)/(p(n-N))$ , and the likelihood ratio in favour of  $\mathbf{R} \in \mathcal{D}$  is given by (cf. (3.6))

$$\frac{P(\mathbf{R} = i)}{pP(\mathbf{R} \in \mathcal{D})} = \frac{1}{Np}.$$

These results are well known; see e.g. [9] and the references therein.

#### 4. SEARCH STRATEGIES

We will now use these results to define strategies to choose a subset of  $\mathcal{D}$  as small as possible, and which contains  $\mathbf{R}$  with a given minimal probability  $\alpha$ .

**4.1. The conditional method.** Let  $\mathcal{D}^k$  be the subset of  $\mathcal{D}$  that corresponds to the  $k$  largest products  $r_i P(\mathbf{R} = i)$  (with some arbitrary rule in case of ties). Furthermore, we let, for  $0 \leq \alpha \leq 1$ ,  $k_{\alpha}$  be the minimal  $k$  for which

$$\sum_{j \in \mathcal{D}^k} r_j P(\mathbf{R} = j) \geq \alpha(r_1 P(\mathbf{R} = 1) + \dots + r_N P(\mathbf{R} = N)),$$

that is,  $k_{\alpha}$  is the smallest  $k$  for which the corresponding sum of the likelihood ratios weighted with the prior probabilities is at least a fraction  $\alpha$  of the total weighted

sum. Finally, we write  $\mathcal{D}^\alpha$  for  $\mathcal{D}^{k_\alpha}$ . Note that in order to determine whether or not  $i \in \mathcal{D}^\alpha$ , one needs the full vector  $\mathbf{r}$ . Note also that  $P(\mathbf{R} \in \mathcal{D}^\alpha)$  depends on the distribution of  $\mathbf{H}_1$  and  $\mathbf{H}_2$ , but  $P(\mathbf{R} \in \mathcal{D}^\alpha \mid \mathbf{LR}_D = \mathbf{r})$  does not (cf. Remark 3.2). The distribution of the cardinality of  $\mathcal{D}^\alpha$  also depends on the distribution of  $\mathbf{H}_1$  and  $\mathbf{H}_2$  (and on  $N$ ).

We now make two observations about the probability that the index  $\mathbf{R}$  is contained in  $\mathcal{D}^\alpha$ . First, in case  $P(\mathbf{R} \in \mathcal{D}) = 1$  it follows from (3.1) that for the unconditional probability  $P(\mathbf{R} \in \mathcal{D}^\alpha)$  we have

$$(4.1) \quad P(\mathbf{R} \in \mathcal{D}^\alpha) \geq \alpha.$$

Secondly, if we do not have  $P(\mathbf{R} \in \mathcal{D}) = 1$ , then from (3.2) we have (for  $i = 1, \dots, N$ ) that

$$\begin{aligned} P(\mathbf{R} = i \mid \mathbf{R} \in \mathcal{D}) &= \sum_{\mathbf{r}} P(\mathbf{LR}_D = \mathbf{r} \mid \mathbf{R} \in \mathcal{D}) P(\mathbf{R} = i \mid \mathbf{R} \in \mathcal{D}, \mathbf{LR}_D = \mathbf{r}) \\ &= \sum_{\mathbf{r}} P(\mathbf{LR}_D = \mathbf{r} \mid \mathbf{R} \in \mathcal{D}) \frac{r_i P(\mathbf{R} = i)}{\sum_{k=1}^N r_k P(\mathbf{R} = k)}. \end{aligned}$$

Hence

$$\begin{aligned} P(\mathbf{R} \in \mathcal{D}^\alpha \mid \mathbf{R} \in \mathcal{D}) &= \sum_{\mathbf{r}} P(\mathbf{LR}_D = \mathbf{r} \mid \mathbf{R} \in \mathcal{D}) P(\mathbf{R} \in \mathcal{D}^\alpha \mid \mathbf{R} \in \mathcal{D}, \mathbf{LR}_D = \mathbf{r}) \\ &= \sum_{\mathbf{r}} P(\mathbf{LR}_D = \mathbf{r} \mid \mathbf{R} \in \mathcal{D}) \sum_{i \in \mathcal{D}^\alpha} \frac{r_i P(\mathbf{R} = i)}{\sum_{k=1}^N r_k P(\mathbf{R} = k)} \\ &\geq \alpha \sum_{\mathbf{r}} P(\mathbf{LR}_D = \mathbf{r} \mid \mathbf{R} \in \mathcal{D}) = \alpha, \end{aligned}$$

where the inequality follows from the definition of  $\mathcal{D}^\alpha$ . The quantity

$$P(\mathbf{R} \in \mathcal{D}^\alpha \mid \mathbf{R} \in \mathcal{D})$$

is called the *efficiency* of  $\mathcal{D}^\alpha$ , the ability to select  $\mathbf{R}$  given that  $\mathbf{R}$  is in the database. We just showed that the efficiency of  $\mathcal{D}^\alpha$  is at least  $\alpha$ .

**4.2. The target-centered method.** For  $0 \leq \alpha \leq 1$ , let  $t_\alpha \geq 0$  be the largest  $t$  for which

$$(4.2) \quad P(\mathbf{LR}(\mathbf{H}_1) \geq t) \geq \alpha.$$

We use these thresholds  $t_\alpha$  to define

$$(4.3) \quad \mathcal{D}_\alpha = \{i \in \mathcal{D} \mid \mathbf{LR}(\mathbf{P}_i) \geq t_\alpha\}.$$

In order to decide whether or not  $i \in \mathcal{D}_\alpha$ , one only needs to know  $r_i$ , and not the full vector  $\mathbf{r}$  as in the case of  $\mathcal{D}^\alpha$ . It follows that

$$P(\mathbf{R} \in \mathcal{D}_\alpha \mid \mathbf{R} \in \mathcal{D}) = P(\mathbf{LR}(\mathbf{H}_1) \geq t_\alpha) \geq \alpha,$$

so also the efficiency of  $\mathcal{D}_\alpha$  is at least  $\alpha$ . In fact, for every  $0 \leq \alpha \leq 1$ ,

$$(4.4) \quad P(\mathbf{R} \in \mathcal{D}_\alpha) \geq \alpha P(\mathbf{R} \in \mathcal{D}).$$

At this point we mention a connection with the criterion proposed in [8] to decide whether or not to further investigate the possibility that the suspect's relative matches a crime stain. In our terminology, they propose a threshold  $s_\beta$  such that

$$\beta = P(\mathbf{LR}(\mathbf{H}_2) \geq s_\beta).$$

If  $N$  is large then (since all but at most one of the database members are distributed as  $H_2$ ) one expects a fraction  $\beta$  of the database to be selected into  $\{i \in \mathcal{D} \mid \text{LR}(P_i) \geq s_\beta\}$ . The relation with  $\mathcal{D}_\alpha$  is as follows. We have

$$\begin{aligned} \alpha &\leq P(\text{LR}(H_1) \geq t_\alpha) \\ &= \sum_{x \geq t_\alpha} P(\text{LR}(H_1) = x) \\ &= \sum_{x \geq t_\alpha} xP(\text{LR}(H_2) = x) \\ &= \sum_{x \geq 0} xP(\text{LR}(H_2) = x \mid \text{LR}(H_2) \geq t_\alpha)P(\text{LR}(H_2) \geq t_\alpha) \\ &= E(\text{LR}(H_2) \mid \text{LR}(H_2) \geq t_\alpha)P(\text{LR}(H_2) \geq t_\alpha). \end{aligned}$$

Now, let  $\beta$  be such that  $t_\alpha = s_\beta$ , then

$$\alpha \leq \beta \cdot E(\text{LR}(H_2) \mid \text{LR}(H_2) \geq s_\beta).$$

Clearly,  $\alpha$  cannot be expressed in  $\beta$  alone but depends also on the target that we are dealing with. It follows that when selecting a database subset according to the threshold  $s_\beta$ , the probability that  $\mathbf{R}$  is selected depends on the specific aspects of the case, whereas for  $\mathcal{D}_\alpha$  (and  $\mathcal{D}^\alpha$ ) it has a uniform lower bound  $\alpha$ .

**4.3. Comparison and interpretation.** We have defined two subsets  $\mathcal{D}^\alpha$  and  $\mathcal{D}_\alpha$ , both with efficiency at least  $\alpha$ . Nevertheless, there are important differences between these approaches that we wish to discuss here.

First of all,  $\mathcal{D}^\alpha$  makes use of the prior probabilities  $P(\mathbf{R} = i)$ , while  $\mathcal{D}_\alpha$  does not. For example, in case of familial searching, geographical information or age could play a role in the definition of prior probabilities  $P(\mathbf{R} = i)$ . Thus,  $\mathcal{D}^\alpha$  uses more information than  $\mathcal{D}_\alpha$ , which seems to give  $\mathcal{D}^\alpha$  an advantage over  $\mathcal{D}_\alpha$ .

There is, however, a reason why the use of  $\mathcal{D}_\alpha$  could be more appropriate in concrete cases. This reason has to do with the interpretation of the probabilities involved, and we explain this next. We can see  $\mathcal{D}^\alpha$  as a random subset of  $\mathcal{D}$  which contains all database members that have yielded likelihood ratios greater than or equal to a *random* threshold. The distribution of this threshold depends on the distributions of both  $H_1$  and  $H_2$  (and on  $N$ , the size of the database). Therefore, a frequentist interpretation requires re-sampling of the database. Indeed, we have defined a subset in such a way that, if we would construct it for many realizations of one copy of  $H_1$  among  $N - 1$  copies of  $H_2$ , a fraction  $P(\mathbf{R} \in \mathcal{D}^\alpha \mid \mathbf{R} \in \mathcal{D})$  of the time we would have included the copy of  $H_1$ .

The interpretation of the probability  $P(\mathbf{R} \in \mathcal{D}_\alpha)$ , on the other hand, is easier. Indeed,  $\mathcal{D}_\alpha$  is a random subset of  $\mathcal{D}$  as well, containing all database members that have yielded likelihood ratios above some threshold, but this time the threshold depends on the distribution of  $H_1$  only (through  $\text{LR}(H_1)$ ). This allows us to make another frequentist interpretation: we choose a realisation of the database (according to  $H_2$ ), and then, keeping the database fixed, repeatedly add one copy of a realisation of  $H_1$ . We can think of  $P(\mathbf{R} \in \mathcal{D}_\alpha \mid \mathbf{R} \in \mathcal{D})$  as the relative frequency of times we would find the special member in  $\mathcal{D}_\alpha$ . This interpretation corresponds well with what one would intuitively understand by the probability of finding the relative since in the forensic practice, the database is (more or less) fixed. From this

point of view it is more appropriate to use  $\mathcal{D}_\alpha$  rather than  $\mathcal{D}^\alpha$  and, importantly, it is also easier to explain to legal representatives what the probabilistic statement really means.

The frequentist considerations above apply to the general framework we have discussed in this paper. In the special case of familial searching however, the drawback of using  $\mathcal{D}^\alpha$  may not be that serious, for the following reason. We explained that for a full frequentist interpretation of  $\mathcal{D}^\alpha$ , one would need to resample the database many times, and that this does not correspond well to legal practice. However, what matters is not so much that we can interpret the full profiles in the database as being resampled, but that we can interpret the *observed likelihood ratios* as being resampled. Suppose that we treat various different familial searching cases (i.e., try to find relatives from various targets) with the same database. Then, when we compute likelihood ratios between the database and a new target, these likelihood ratios depend on the newly sampled target profile, and it is to be expected that for an independent sequence of target profiles, the observed likelihood ratios corresponding to the fixed profiles in the database are more or less independent. To test this, in the next section we investigate using computer simulation to what extent the frequentist interpretation that we have for  $\mathcal{D}_\alpha$  is valid for  $\mathcal{D}^\alpha$  as well. That is, we draw many targets independently according to  $H_2$  (i.e., at random using population allele frequencies), and add their simulated relatives to the same database. We see how many of these relatives are found in  $\mathcal{D}^\alpha$ , on average over all targets. This we will compare to adding many relatives of the *same* target to resampled databases.

Finally, we mention the fact that when the database is large and uniform priors are used, the sets  $\mathcal{D}^\alpha$  and  $\mathcal{D}_\alpha$  will be very similar. This is due to the fact that the law of large numbers implies that the sum of the likelihoods above  $t_\alpha$  divided by  $N$  will be close to  $\alpha$ . Hence the random threshold associated with  $\mathcal{D}^\alpha$  (discussed above) will with very high probability be very close to  $t_\alpha$ . This argument can be made precise in the form of a limit statement in probability or almost surely.

## 5. FAMILIAL SEARCHING IN THE DUTCH NATIONAL DNA DATABASE

**5.1. Methods and notation.** All our simulations were programmed in-house with Mathematica software. We let  $\mathcal{D}_{NL}$  be the Dutch National DNA Database (as per mid 2010, all duplicate profiles removed and only considering the  $N = 99,979$  profiles for which all ten SGM+ loci were typed). Allelic ladders and allele frequencies were taken from  $\mathcal{D}_{NL}$ , and mutation rates were based on those published by the NIST<sup>1</sup>.

According to these allele frequencies, target profiles  $C_1, \dots, C_{100}$  were sampled (pseudo)randomly to serve as the targets whose relatives we want to find using familial searching. For each of these target profiles we sampled 50,000 children and 50,000 siblings, using mutation probabilities as described above. Then we computed the likelihood ratios in favour of paternity (the *Paternity Index PI*) between the  $C_i$  and their children, and those in favour of siblingship (the *Sibling Index SI*) between the  $C_i$  and their siblings. These allow us to estimate the thresholds  $t_\alpha$  (cf. (4.2)) for the paternity and sibling cases. Both *PI* and *SI* were computed taking

---

<sup>1</sup><http://www.cstl.nist.gov/strbase/mutation.htm>, based on data compiled by the American Association of Blood Banks.

the possibility of mutation into account, with the same mutation model (a stepwise one) as for the generation of the relatives.

The DNA profiles in  $\mathcal{D}_{NL}$  are labeled  $d_1, \dots, d_N$ ; they can be viewed as a sample of independent copies of  $\mathbb{H}_2$  that is fixed throughout. By  $PI(C_i, d_j)$  we mean the Paternity Index between the target profile  $C_i$  and database profile  $d_j$ . Thus,  $PI(C_i, d_j)$  can, for each target separately, be interpreted as a realization  $r_j$  of the random variables  $LR(\mathbb{H}_2)$  in the preceding sections. Notation for the sibling case is similar. We will sometimes write  $KI$ , for *Kinship Index*, when we mean that the discussion holds for both  $PI$  and  $SI$ .

Finally, we have also computed the random match probability (RMP) of each target profile. On a locus with alleles  $(a, b)$ , the RMP is equal to  $p_a p_b (2 - \delta_{a,b})$  where  $p_i$  is the allele frequency of allele  $i$  and  $\delta_{a,b} = 0$  if  $a \neq b$  and  $\delta_{a,a} = 1$ . The RMP of a DNA profile is then the product over all involved loci, since we assume all loci to be independent.

**5.2. Total likelihood ratio with the database.** For all 100 targets  $C_i$ , we computed the sums  $|KI(C_i, \mathcal{D}_{NL})| = \sum_{k=1}^N KI(C_i, d_k)$ . The mean  $|PI(C_i, \mathcal{D}_{NL})|$  was 102,200 (with sample standard deviation 94,500), the mean  $|SI(C_i, \mathcal{D}_{NL})|$  was 93,500 (with sample standard deviation 42,200). These results seem consistent with what we expect from Proposition 2.2. Indeed, since all target profiles were randomly generated, they do not have a true relative in the database, and hence  $E(|KI(C_i, \mathcal{D}_{NL})|) = N$  according to Proposition 2.2.

**5.3. The conditional method in the Dutch National DNA Database.** We have investigated (by simulation) what the probability is that a relative of a fixed target is found in  $\tilde{\mathcal{D}}_{NL}^\alpha$ , where  $\tilde{\mathcal{D}}_{NL}$  is the extension of  $\mathcal{D}_{NL}$  with a relative of the considered target. We take a uniform prior distribution of  $\mathbf{R}$  on  $\tilde{\mathcal{D}}_{NL}$ .

To do so, we have simulated relatives  $R_{i,j}$  ( $i = 1, \dots, 100; j = 1, \dots, 500$ ) of each type (children and siblings), where  $R_{i,j}$  is a relative of target profile  $C_i$ . For each relative  $R_{i,j}$ , define its *rank* to be equal to  $k$  if and only if there are exactly  $k - 1$  database members that have a greater kinship index with  $C_i$  than  $R_{i,j}$ . We also define

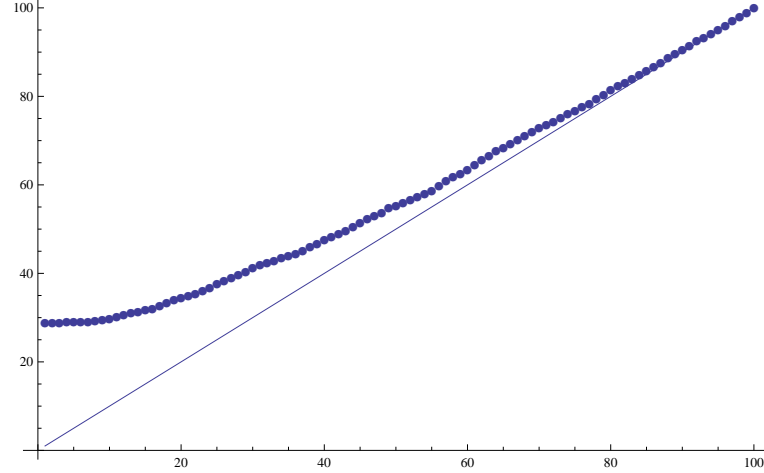
$$t_{i,j} = \frac{\sum_{x: KI(C_i, d_x) > KI(C_i, R_{i,j})} KI(C_i, d_x)}{KI(C_i, d_1) + \dots + KI(C_i, d_N) + KI(C_i, R_{i,j})}.$$

Assuming a uniform prior of  $\mathbf{R}$  on  $\tilde{\mathcal{D}}_{NL}$ ,  $t_{i,j}$  is the greatest  $t \geq 0$  such that  $R_{i,j} \notin \tilde{\mathcal{D}}_{NL}^t$ . Thus,  $R_{i,j} \in \tilde{\mathcal{D}}_{NL}^\alpha$  if and only if  $\alpha > t_{i,j}$ .

For each  $C_i$  and for  $\alpha \in \{0.01, \dots, 0.99, 1\}$ , we have compared  $\alpha$  to the fraction  $\beta_{i,\alpha}$  of  $t_{i,j}$  that are smaller than  $\alpha$ ; this fraction  $\beta_{i,\alpha}$  is the observed probability for relatives of  $C_i$  to be in  $\tilde{\mathcal{D}}_{NL}^\alpha$ . Finally, we have also computed  $\beta_\alpha$  as the average over all  $\beta_{i,\alpha}$ . Thus,  $\beta_\alpha$  estimates the probability that if one adds a relative  $\mathbf{R}$  of a random target profile to *this* database  $\mathcal{D}_{NL}$ , that  $\mathbf{R}$  is in  $\tilde{\mathcal{D}}_{NL}^\alpha$ .

The probability that the relative of a target  $C$  is in  $\tilde{\mathcal{D}}_{NL}^\alpha$  is called the *probability of detection* (POD) for  $C$  in the Dutch National Database  $\mathcal{D}_{NL}$ . The number  $\beta_\alpha$  therefore gives an estimate of the average (over all targets) probabilities of detection POD. Note that a POD is only defined in connection to a fixed database, in this case  $\mathcal{D}_{NL}$ .

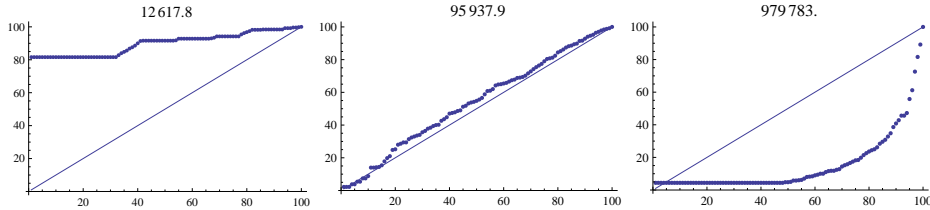
FIGURE 1. The average POD as a function of  $\alpha$ , averaged over 100 target profiles  $C$ , Paternity Index.



5.3.1. *Paternity Index.* For paternity indices, the result of our simulations is displayed in Figure 1. For all  $\alpha$  the average POD of  $\tilde{\mathcal{D}}_{NL}^\alpha$  is at least  $\alpha$ . For small  $\alpha$ , it exceeds  $\alpha$  substantially and as  $\alpha$  increases, the average POD of  $\tilde{\mathcal{D}}_{NL}^\alpha$  approaches  $\alpha$ . This is a consequence of the definition of  $\tilde{\mathcal{D}}_{NL}^\alpha$  as being a subset that contains the  $k$  greatest  $PI$  for some  $k$ . As  $\alpha$  increases and  $\tilde{\mathcal{D}}_{NL}^\alpha$  becomes larger, we add individuals with smaller  $PI$ , and we expect  $\beta_\alpha$  to become closer to  $\alpha$ . However, substantial variation between target profiles is to be expected, since the presence or absence of database members that have a large  $PI$  with the target by chance will affect all  $\tilde{\mathcal{D}}_{NL}^\alpha$  with this target.

Indeed, the observed variation between target profiles was substantial. We highlight three very different results in Figure 2. The number above each graph is

FIGURE 2. The POD as a function of  $\alpha$ , for three target profiles.



$|PI(C_i, \mathcal{D}_{NL})|$ , the total Paternity Index with the Dutch National DNA Database (without the relative). We make a few remarks about these results:

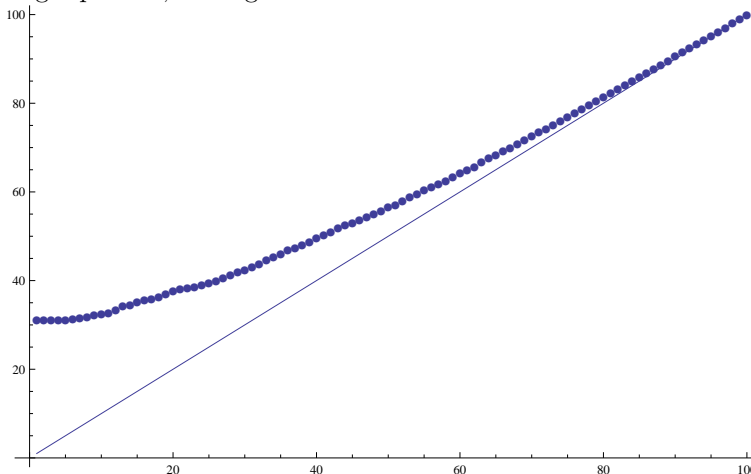
- (1) The first  $|PI(C, \mathcal{D}_{NL})|$  is much smaller than expected. For this target  $C$ ,  $\max_j PI(C, d_j) = 6,000$ . The probability that  $PI(C, R)$  is at least 6,000 is equal to 0.82. This is what we find as probability of detection for small  $\alpha$ , as we can see in the figure.
- (2) The second graph corresponds to a target for which the total  $PI$  with the database was almost exactly as expected. Moreover, this total  $PI$  is realized

as the sum of many likelihood ratios of comparable magnitude. The largest of these is equal to 3,100. Only 2% of children of  $C$  have a  $PI$  with  $C$  of at least 3,100. Therefore the probability that the true relative has a higher  $PI$  with  $C$  than all the unrelated individuals in  $\mathcal{D}$  is only 0.02.

- (3) The third graph corresponds to a profile  $C$  for which  $|PI(C, \mathcal{D}_{NL})| = 979,000$ . Almost all of this is accumulated on one database member  $d$  with  $PI(C, d) = 917,000$ . The probability that  $PI(C, R)$  exceeds this is 0.044. In the graph, the probability that the relative of  $C$  is found in  $\mathcal{D}^\alpha$  is also equal to 0.044, for  $\alpha$  up to about 0.5.

5.3.2. *Sibling Index.* For sibling indices, the result of our simulations is displayed in Figure 3. In this case as well, the variation between different target profiles was substantial.

FIGURE 3. The average POD as a function of  $\alpha$ , average over 100 target profiles, Sibling Index.

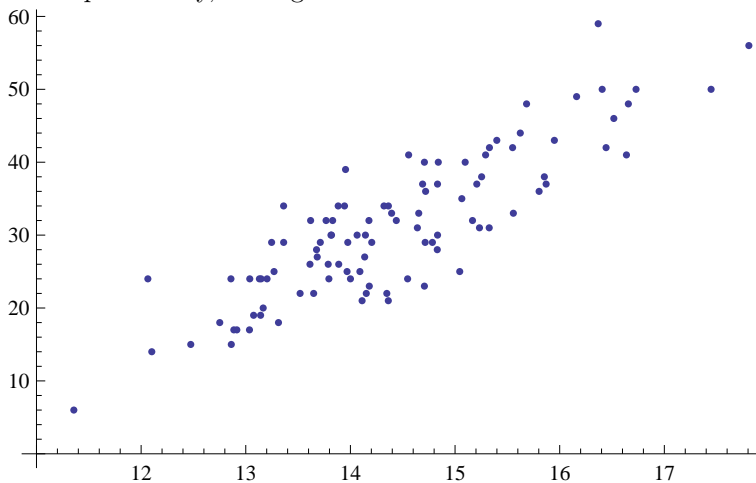


5.4. **The conditional method in resampled databases.** In this section we compare the above estimates of the probabilities of detection with estimates of the efficiency of the conditional method in resampled databases of roughly the same size as  $\mathcal{D}_{NL}$ . To do so we have, for each target profile  $C_i$  as above, simulated 100 relatives  $R'_{i,j}$  and databases  $\mathcal{D}_{i,j}$  with  $N = 100,000$ . As in the previous section, we have determined  $t_{i,j}$  as the largest  $\alpha$  such that  $R'_{i,j} \notin \tilde{\mathcal{D}}_{i,j}^\alpha$ , and used these numbers to determine  $\beta'_{i,\alpha}$  and  $\beta'_\alpha$  whose definitions are analogous to their earlier counterparts. Since these simulations are very time consuming, we have only carried them out for siblings.

5.4.1. *Observed efficiency.* As was to be expected, the graphs of  $\beta'_{i,\alpha}$  as a function of  $\alpha$  differ much less between different targets than in the fixed DNA database above: every target displays the behaviour that for small  $\alpha$  the observed efficiency  $\beta'_{i,\alpha}$  is greater than  $\alpha$  (in fact,  $\beta'_{i,\alpha}$  lies between 0.06 and 0.59 for  $\alpha = 0.01$ ) and as  $\alpha$  increases, the  $\beta'_{i,\alpha}$  approach  $\alpha$ . There is, however, still some variation. This is, at least in part, due to the fact that the distribution  $H_1$  is different for different

targets, meaning that some targets tend to have higher sibling indices with their siblings than others, and this results in a greater efficiency of  $\mathcal{D}^\alpha$  for small  $\alpha$ . In Figure 4, we plot  $\beta'_{i,0.01}$  as a function of  $-\text{Log}_{10}$  of the random match probability of target profile  $C_i$ . From the figure we see that the rarer the target profile, the greater  $\beta'_{i,0.01}$  tends to be.

FIGURE 4. Observed efficiency  $\beta'_{i,0.01}$  as a function of the random match probability, Sibling Index.



The observed overall efficiency  $\beta'_\alpha$  is extremely close to the observed probabilities of detection  $\beta_\alpha$  displayed in Figure 3. In fact, the difference between the  $\beta'_\alpha$  of this section (the average efficiency) and the  $\beta_\alpha$  in Figure 3 (average probability of detection) is on average over  $\alpha \in \{0.01, \dots, 0.99, 1\}$  equal to  $-0.0022$  and never greater (in absolute value) than  $0.0084$ .

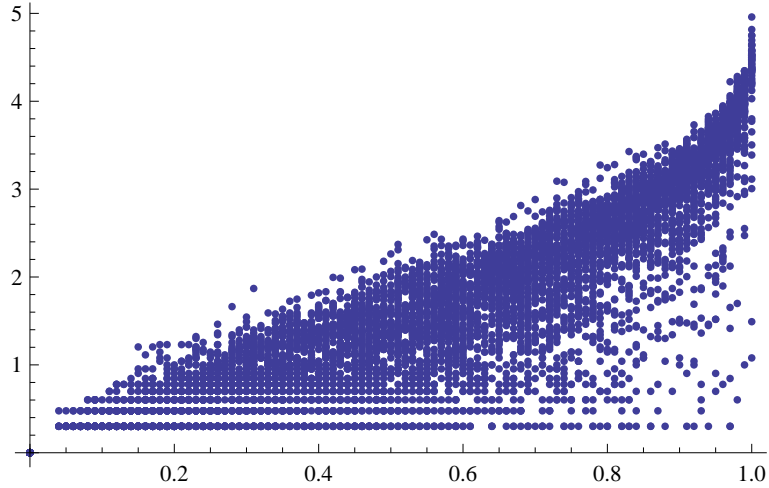
5.4.2. *Observed ranking and size of  $\tilde{\mathcal{D}}^\alpha$ .* We have also, for each  $R_{i,j}$ , computed its rank  $k_{i,j}$ . The rank of  $R_{i,j}$ , defined in Section 5.3, is the size of the smallest  $\tilde{\mathcal{D}}^\alpha_{i,j}$  that includes  $R_{i,j}$ . We plot the observed results in Figure 5, where  $k_{i,j}$  is plotted as a function of  $t_{i,j}$ . Several things can be read off this graph.

First, when one looks at a constant  $t_{i,j}$ , one sees the rank in which relatives are found that are detected by  $\mathcal{D}^\alpha$  with  $\alpha > t_{i,j}$ . For example, with  $t_{i,j} = 0.5$ , we see from the figure that this corresponds to  $k_{i,j} \leq 100$ . Thus, in all cases where the database contains a relative and  $\alpha$  is set at 0.5, the relative is found in about 55% of the cases (as can be seen from Figure 3), and in the cases where it is found, it occupies a position between 1 and 100 when the database is ordered according to decreasing likelihood ratio.

Second, when we keep  $k_{i,j}$  fixed one can see what  $\alpha$  is typically needed to find the relative if it is ranked in position  $k_{i,j} + 1$ . For example, one sees that  $k_{i,j} = 1,000$  corresponds to  $t_{i,j}$  of 0.8 and more, meaning that in all cases where the relative is ranked around position 1,000, an efficiency of at least 0.8 is usually required to find it.

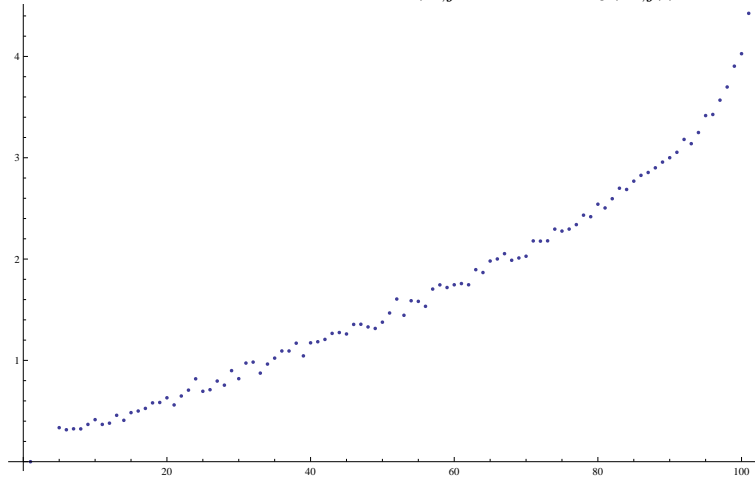
A summary of this graph is plotted in Figure 6, where we plot for each  $\alpha \in \{0, 0.01, \dots, 0.99\}$ , the average rank of relatives for which  $t_{i,j}$  is nearest to  $\alpha$ . This

FIGURE 5. Observed pairs  $(t_{i,j}, \text{Log}_{10}(k_{i,j}))$



gives the average rank of a relative that would be found in  $\mathcal{D}^\alpha$ , but not in  $\mathcal{D}^\beta$  for  $\beta < \alpha$ .

FIGURE 6. Observed pairs  $(t_{i,j}, \text{Mean Log}_{10}(k_{i,j}))$



We will compare this below to the average size of  $\mathcal{D}_\alpha$ .

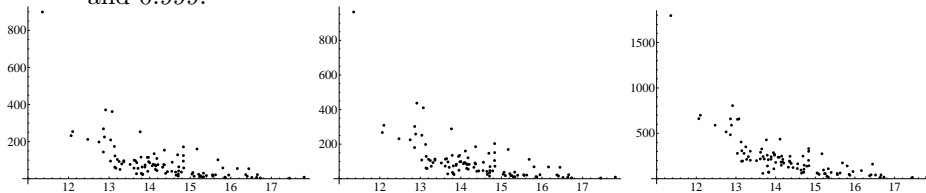
5.5. **The size of  $\mathcal{D}_\alpha$ .** Recall that we have, for each of the target profiles  $C_i$ , simulated 50,000 children and 50,000 siblings, in order to estimate the profile-dependent thresholds  $t_\alpha$  needed for  $\mathcal{D}_\alpha$ .

5.5.1. *Paternity Index.* Mutation frequencies are such that about 99% of parent-child pairs share an allele on all ten SGM+ loci. Since a mutation is rare, it has a strong effect on the *PI* and therefore the candidates selected by a search for database members who share an allele on each locus will be almost the same

as those selected by  $D_\alpha$  with  $\alpha = 0.99$ . This yields, on average over all  $C_i$ , 96 candidates from  $\mathcal{D}_{NL}$ .

Thus, the  $\mathcal{D}_\alpha$  approach is most interesting for  $\alpha$  at least equal to this. We have therefore determined the 1%, 0.5% and 0.1% quantiles of the observed  $PI$  between  $C_i$  and its children and used these as thresholds  $t_\alpha$  for  $\alpha \in \{0.99, 0.995, 0.999\}$ . The resulting cardinalities of  $\mathcal{D}_{NL,\alpha}$  are plotted in Figure 7 below as a function of  $-\log_{10}(\text{RMP})$  of the target profile.

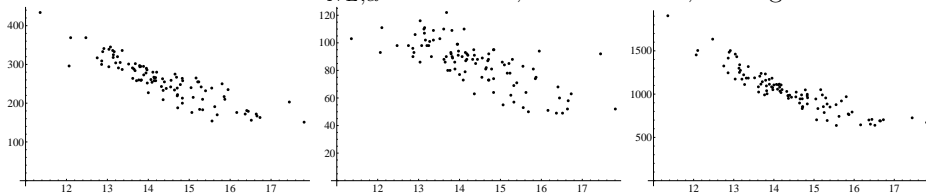
FIGURE 7. Size of  $\mathcal{D}_{NL,\alpha}$  for parent-child relation;  $\alpha=0.99, 0.995$  and  $0.999$ .



The mean size of  $\mathcal{D}_{NL,\alpha}$  was equal to 99, 102, resp. 226, with substantial variation. For a rare the profile, fewer candidates are in  $\mathcal{D}_{NL,\alpha}$ . This indicates that, for a database of the size of  $\mathcal{D}_{NL}$ , it is for most targets possible to extract a parent or a child with very high probability: even for  $\alpha = 0.999$ , the size of  $\mathcal{D}_\alpha$  is usually such that it should not be too unpractical to further investigate all these database members (e.g., by determining more autosomal loci or Y-chromosomal DNA profiles).

5.5.2. *Sibling Index*. Naturally, finding a sibling in a database is harder than finding a parent or child. We have determined the  $t_\alpha$  for  $\alpha$  equal to 0.70, 0.80 and 0.90. The resulting sizes of  $\mathcal{D}_{NL,\alpha}$  are plotted in Figure 8. The mean sizes are 85, 258, 1038 respectively.

FIGURE 8. Size of  $\mathcal{D}_{NL,\alpha}$  for  $\alpha=0.70, 0.80$  and  $0.90$ , Sibling Index



The size of  $\mathcal{D}_{NL,\alpha}$  increases rapidly with  $\alpha$ . Although even for 0.90, upgrading these DNA-profiles to 15 autosomal loci and Y-chromosomal profiles would downsize  $\mathcal{D}_{NL,\alpha}$  to more manageable numbers, it seems impractical to do this in reality. In addition, the threshold on the likelihood ratio is so low that it would result in the further (genetic, perhaps even tactical) investigation of people based on a very weak indication, perhaps even a counter-indication, for an a priori extremely unlikely event (being a sibling of the target).

**5.6. Comparison of simulations.** The simulations indicate that, even though there is a conceptual difference between probability of detection and efficiency, the average probability of detection in the Dutch National DNA Database coincides with the average efficiency of  $\mathcal{D}^\alpha$  for databases of the same size, cf. Section 5.4.1. In other words, the average probability of detection in  $\mathcal{D}_{NL}$ , averaged over all targets, is the same as the average efficiency of  $\mathcal{D}^\alpha$  (averaged over all targets) for a database  $\mathcal{D}$  of the same size as  $\mathcal{D}_{NL}$ . Therefore, even if we (as we do in practice) do not resample the database but look for relatives in the same database for all targets, then the probability of finding a relative in  $\mathcal{D}^\alpha$  when database and relative are resampled is the same as the long term success rate of finding the relative of varying targets in  $\mathcal{D}^\alpha$  while the database is kept constant. On the other hand, for  $\mathcal{D}_\alpha$ , the interpretation of the efficiency  $P(\mathbf{R} \in \mathcal{D}_\alpha \mid \mathbf{R} \in \mathcal{D})$  does not require resampling of the database, hence also holds in a fixed one. This makes the  $\mathcal{D}_\alpha$  method easier to interpret.

Finally, we note that although the efficiency of  $\mathcal{D}^\alpha$  does not really depend on the target (it is approximately  $\alpha$  for all targets and sufficiently large  $\alpha$ ), the cardinality of  $\mathcal{D}^\alpha$  does depend on the target. For example, a familial search for parents and children will yield smaller  $\mathcal{D}^\alpha$  than one for siblings. But also if the type of relative is fixed, differences are to be expected between different targets: a target profile with many rare alleles will tend to have a greater kinship index with its relative than a target with only common alleles.

#### REFERENCES

- [1] F.R. Bieber, C.H. Brenner, and D. Lazer, *Finding Criminals Through DNA of Their Relatives*, Science **312** (2006), 1315–6.
- [2] J.M. Curran and J.S. Buckleton, *Effectiveness of familial searches*, Science and Justice **84** (2008), 164–7.
- [3] J. Ge, R. Chakraborty, A. Eisenberg, et al., *Comparisons of Familial DNA Database Searching Strategies*, J. For. Sc. **in press** (2011).
- [4] T. Hicks, F. Taroni, J. Curran, et al., *Use of DNA profiles for investigation using a simulated national DNA database: Part II. Statistical and ethical considerations on familial searching*, For. Sc. Int.:Genetics **4** (2010), no. 5, 232–8.
- [5] G. Miller, *Familial DNA Testing Scores a Win in Serial Killer Case*, Science **329** (2010), 262.
- [6] S. Myers et al., *Searching for first-degree familial relationships in california’s offender DNA database: Validation of a likelihood ratio-based approach*, For. Sc. Int.: Gen. **5** (2011), no. 5, 493–500.
- [7] Scientific Working Group on DNA Analysis Methods Ad Hoc Committee on Partial Matches, *SWGDM Recommendations to the FBI Director on the “Interim Plan for the Release of Information in the Event of a ‘Partial Match’ at NDIS”*, For. Sc. Comm. **11** (2009), no. 4.
- [8] M. Sjerps and A.D. Kloosterman, *On the consequences of DNA profile mismatches for close relatives of the suspect*, Int. J. Legal Med. **112** (1999), 176–180.
- [9] K. Slooten and R. Meester, *Forensic Identification: the Island Problem and its generalizations*, Stat. Neerl. ? (2011), ?

NETHERLANDS FORENSIC INSTITUTE, P.O. BOX 24044, 2490 AA THE HAGUE, THE NETHERLANDS

*E-mail address:* k.slooten@nfi.minjus.nl

VU UNIVERSITY AMSTERDAM, DE BOELELAAN 1081, 1081 HV AMSTERDAM, THE NETHERLANDS

*E-mail address:* rmeester@few.vu.nl