Forensic Science International: Genetics 13 (2014) 90–103

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Optimal strategies for familial searching

Maarten Kruijver [a,*], Ronald Meester [a], Klaas Slooten [a,b]

[a] Department of Mathematics, VU University, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[b] Netherlands Forensic Institute, P.O. Box 24044, 2490 The Hague, The Netherlands

ARTICLE INFO

Article history:
Received 9 January 2014
Accepted 11 June 2014

ABSTRACT

Familial searching is the process of finding potential relatives of the donor of a crime scene profile in a DNA database. Several authors have proposed strategies for generating candidate lists of potential relatives. This paper reviews four strategies and investigates theoretical properties as well as empirical behavior, using a comprehensive simulation study on mock databases. The effectiveness of a familial search is shown to highly depend on the case profile as well as on the tuning parameters. We give recommendations for proceeding in an optimal way and on how to choose tuning parameters both in general and on a case-by-case basis. Additionally we treat searching heterogeneous databases (not all profiles comprise the same loci) and composite searching for multiple types of kinship. An R-package for reproducing results in a particular case is released to help decision-making in familial searching.

1. Introduction

Many countries maintain DNA databases containing profiles of convicted offenders (some also include suspects), which are routinely compared to crime scene DNA profiles. At a much smaller scale these databases are used for identifying close relatives of potential suspects through a process called familial searching [1]. Searches typically proceed by computing *likelihood ratios* (*LR*s) comparing the hypotheses of unrelatedness, full siblings and parent/offspring for the sample with all the database members based on their autosomal DNA profiles. These likelihood ratios are usually called *kinship indices* (*KI*s) in this context. A familial search strategy generates a candidate list of database members with the highest *KI*s. Any strategy strikes a balance between the probability of including a true relative on the one hand and the length of the candidate list on the other. A follow-up procedure consisting of additional tests such as typing additional autosomal loci or Y-STR/mitochondrial profiling excludes many candidates so that ultimately very few false leads remain. Generating candidate lists efficiently is key to effective familial searching.

Several authors have proposed strategies for generating candidate lists for use in further investigations [2–4]. The purpose of the current paper is first, to evaluate the state of the art in familial search strategies, and second, to investigate how the effectiveness and workload of a familial search are affected by the number of loci typed, the database size and the search strategy chosen. We study theoretical properties of four familial search strategies and give recommendations on how to choose a strategy and tuning parameters, both in general and on a case-by-case basis. Additionally we provide practical guidelines on how to deal with databases for which not all persons are profiled on the same loci. Finally, we treat the problem of how to perform an optimal search for multiple types of kinship, for various optimality criteria.

The effectiveness of familial searching has been investigated by a number of authors, mostly using simulation studies concerned with either determining the *KI*-rank of true relatives in a database or the estimation of exceedance probabilities of *KI*-thresholds. Bieber et al. [1] demonstrate the potential of familial searching and determine the empirical *KI*-rank of simulated true full siblings and parents/offspring in a simulated database. Curran and Buckleton [5] undertake a similar study but search the New Zealand DNA database instead of a simulated one. Hicks et al. [6] determine the empirical rank of simulated relatives and evaluate *KI*-thresholds in a simulated database incorporating population substructure. Additionally Hicks et al. pay attention to the ethical aspects of familial searching. Ge et al. [2] evaluate *KI*-thresholds for simulated pairs of relatives and unrelated persons. Actual sibling pairs are used by Reid et al. [7], who determine the distribution of their *KI*-rank in a simulated database. A detailed introduction to familial searching containing a literature review dated 2010 can be found in [8]. More recently, the experiences with familial searching and modus operandi are documented for the state of California [3] and

* Corresponding author. Tel.: +31 633850120.
E-mail addresses: m.v.kruijver@vu.nl (M. Kruijver), r.w.j.meester@vu.nl (R. Meester), k.slooten@nfi.minvenj.nl (K. Slooten).

http://dx.doi.org/10.1016/j.fsigen.2014.06.010
1872-4973/

the United Kingdom [9]. The strategy in use in California is evaluated using a simulation study by Rohlfs et al. [10], who emphasize the possibility of generating false leads resulting from the presence of distant relatives.

In the current paper, we take an approach that differs from the simulation studies mentioned above in a number of ways. The focus on comparing different search strategies rather than showing the utility of familial searching is new to the literature. We also give new theoretical results on optimality of search strategies. Moreover the treatment of optimal searching in heterogeneous (sub-)databases and composite searching are new to the forensic literature. Another difference with existing studies is that we emphasize that in familial searching the database remains fixed, while no case profile is the same. Therefore in our empirical study, we consider a high number of sampled case profiles and determine the behavior of different search strategies in a fixed database. We emphasize that stochastic properties of the search largely depend on the case profile under consideration. Furthermore we offer an R-package with the possibility to study familial search strategies with ease and encourage the reader to do so using parameters of their interest.

The structure of this paper is as follows. Four search strategies are introduced in Section 2. These strategies are evaluated theoretically in Section 3, while Section 4 presents a simulation study illustrating the effects on the length of the candidate list and detection probability of database size, the characteristics of the particular case profile and the multiplex used. Recommendations are given on how to apply the strategies in general and on a case-by-case basis in Section 5. Section 6 treats the case of heterogeneous databases (consisting of profiles obtained using different multiplexes) and Section 7 treats composite searching for multiple types of kinship. Finally, Section 8 contains the main conclusions in short.

2. Four search strategies

Suppose a stain is found at a crime scene and is believed to have been left by an unidentified perpetrator, whom we refer to as C. The multi-locus genotype of this person is denoted G_C. Potential relatives of this person are sought in a database $\mathcal{D}$ of size N. For each database member, we evaluate hypotheses concerning the relation between the database member and the unidentified offender C. We consider the hypotheses unrelatedness, a full sibling relationship and a parent/offspring relationship. For database member i with (multi-locus) genotype G_i, the *LR* for being full siblings is given by

$$SI = \frac{\mathrm{P}(G_C, G_i | \text{ full siblings})}{\mathrm{P}(G_C, G_i | \text{ unrelated})}, \tag{1}$$

which is known as the *sibling index* (*SI*). Similarly, the *LR* comparing the hypotheses of a parent/offspring relationship and unrelatedness is known as the *parent index* (*PI*). Computing the *SI* or *PI* requires a population genetic model. We adopt a simple model ignoring the effects of mutation, dropout, linkage and population (sub)-structure. While these effects might be important in practice, including them here would add little to the present exposition and reduce comparability with other studies. Our computations make use of allelic frequencies as they occur in the reference database (containing DNA profiles of 2085 males) from the Netherlands Forensic Institute. For now we assume that a familial search is conducted for one type of relative (i.e., full sibling or parent/offspring) and that the database consists solely of full profiles comprising a single set of loci. We introduce four strategies for generating candidate lists.

Top-k. The first strategy selects the database members making up the top-k of *KI*s. The largest *KI*s give the strongest indication for relatedness, so working down the list of *KI*s means that the most promising profiles are treated first. Not many authors have explicitly discussed the use of the top-k strategy. Nevertheless most studies on familial searching determine the empirical rank of a true relative using simulation studies [1,11,5] and provide an estimate of the probability by which a true relative would be found in a top-k. Myers et al. [3] report that the state of California implements a strategy in which a candidate list of 168 profiles is used in a follow-up procedure of Y-STR typing. The choice for $k = 168$ is motivated by the practical consideration that two plates of 84 samples can be used in the follow-up.

Fixed KI-threshold. The second strategy selects all database members with a *KI* above some fixed threshold. That is, for some fixed threshold t (e.g., 1000 or 10,000), select all database members for which $KI > t$. In contrast with the top-k strategy, the number of profiles that is selected is not fixed when implementing a *KI*-threshold. This strategy is discussed by Ge et al. [2], who conduct a simulation study to estimate the probability with which a true relative would be selected for various *KI*-thresholds.

Profile-centered. The third strategy also employs a *KI*-threshold, but it chooses this threshold such that a true relative would be selected with some specified probability $0 < \alpha \leq 1$. This is achieved by choosing the threshold as the $(1-\alpha)$-quantile of the distribution of the *KI* for a relative of C. The threshold depends on the case profile under consideration, which is why Slooten and Meester call this strategy the profile-centered method [4].

Conditional. The fourth strategy also appears in [4]. It is based on posterior probabilities of relatedness obtained from prior probabilities and the *KI*s. Assume prior probabilities of relatedness $0 \leq \pi_i \leq 1$ for the database members $i = 1, \ldots, N$. Furthermore, set $\pi_D = \sum_{k=1}^{N} \pi_k$, which is understood as the prior probability that the database contains a relative, and set $\pi_0 = 1 - \pi_D$. Slooten and Meester derive that the posterior probability p_i of relatedness for database member i equals:

$$p_i = \mathrm{P}(i \text{ is the relative} \,|\, KI_1, \ldots, KI_N) = \frac{KI_i \pi_i}{\sum_{k=0}^{N} KI_k \pi_k},$$

under the assumption that there is at most one true relative in the database and setting $KI_0 = 1$. Note that the formula evaluates the posterior probability of relatedness for database member i, *conditional* on all likelihood ratios, which explains the name of the method. A subset of profiles with posterior probability at least α is found by picking the database members with the largest p_is such that their sum is greater than or equal to α. Note that such a subset need not exist; in the extreme case where $\pi_0 = 1$, all $p_i = 0$. It does, however, exist when $\pi_0 = 0$ (corresponding to the case where it is certain that the database contains exactly one relative). The p_i that one obtains for $\pi_0 = 0$ are equal to

$$p_i = \frac{KI_i \pi_i}{\sum_{k=1}^{N} KI_k \pi_k},$$

and consequently, if one chooses a subset such that the sum of these p_is is at least α, then it will contain the relative, which is now known to be present somewhere within the database, with probability at least equal to α. The conditional method, contrary to the others, can be interpreted as a Bayesian method which allows one to make posterior probability statements for each individual in the database. One can use π_D and the observed *KI*s to obtain the posterior probability that the database contains a relative. The likelihood ratio for this (which only makes sense for $0 < \pi_D < 1$) is given by $\sum_{i=1}^{N} KI_i(\pi_i/\pi_D)$. So, contrary to the other methods, the search results can be used to calculate how strong the indication is that the database does contain a relative.