Familial database search on two-person mixture

Yuk-Ka Chung, Wing K. Fung*, Yue-Qing Hu

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China

**ABSTRACT**

This work presents how the familial database search can be applied to forensic investigations in which the DNA evidence is a mixture. The forensic identification is based on a ranked list of the likelihood ratios for each candidate in the DNA database of being the relative of the contributor of the forensic sample. A simulation study using Hong Kong Chinese data demonstrated the comparable effectiveness of our approach that in 90% of the cases, the sibling of the contributor can be found in the top 50 profiles, and the parent/child can be found in the top 10 profiles. We also discuss the prediction on the performance of the search that can guide the decision of the police force to initiate the investigation.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

In the context of forensic science, a deoxyribonucleic acid (DNA) profile refers to a set of numbers representing the genetic characteristics of an individual recorded for the purpose of identification. Since its introduction by Jeffreys et al. (1985), DNA profiling, also known as the DNA fingerprinting, has become a powerful technique for human identification. As discussed in detail elsewhere (Gastwirth, 2000; Chakraborty et al., 1999), short tandem repeat (STR) typing can provide sufficient information and adequate statistical strength for forensic identification and is the most prevalent DNA profiling method used in the forensic community. Many countries have built up their offender DNA databases on the basis of 10 to 15 STR loci. For example, as of April 2009, there are about 7 million offender profiles in the US Federal DNA database CODIS (http://www.fbi.gov/hq/lab/html/codis1.htm) and about 30 to 40 thousand profiles in the Hong Kong DNA database (http://www.govtlab.gov.hk/english/abt_fsd_bss.htm). In recent years, DNA database searching that compares DNA profiles of a trace sample and the individuals in the database has played an important role in forensic identification. In crime investigation, the police office can open an investigation on those individuals whose DNA profiles are perfectly matched with the trace sample found at the crime scene. See Balding and Donnelly (1995), Stockmarr (1999) and Meester and Sjerps (2003), among many others, for comprehensive discussions on the evaluation of the evidentiary value of perfect matches from DNA database search.

When no perfect match is found between the profiles of the trace sample and all the profiles in the database, the partial matches can still provide information to guide the detectives for identifying individuals who appear to be relatives of the perpetrator. Over the years, familial search has been shown to be an effective forensic tool that can increase the number of suspects identified through DNA evidence. A common approach to familial search is to rank the individuals in the database based on the number of alleles they shared with the crime trace or the likelihood ratio (LR) of the following hypotheses:

\[ H_{p1}: \] a relative of the individual is a contributor to the crime trace;

\[ H_{d1}: \] a relative of the individual is not a contributor to the crime trace.
The top-listed individuals will then become the candidates for further investigations. The formulae for calculating LR for various two-person relationships were provided by Lee et al. (2001). The effectiveness of this approach has been evaluated in various articles by using simulated and real DNA databases (Cavallini and Corradi, 2006; Bieber et al., 2006; Curran and Buckleton, 2008; Cowen and Thomson, 2008; Reid et al., 2008). The LR approach has been demonstrated to perform better than the allele sharing approach, with at least 70% chance to detect the true sibling and parent/child in 100 top-listed profiles.

In practical crime cases, it is not uncommon that the biological trace found at the crime scene contains DNA profiles from more than one person. For probable cause case in which the suspect is identified by non-DNA evidence, the evaluation of the DNA mixture has been studied extensively (Weir et al., 1997; Fukhansky and Bär, 1998; Curran, 1999; Fung and Hu, 2000). Some contributions have been made for the situation when the suspect is not typed but his/her relatives are (Fukhansky and Bär, 2000; Hu and Fung, 2003, 2005). Recently, Chung et al. (2009) discussed how the forensic evidence of DNA mixtures can be evaluated when the suspect is identified through database search and derived general formulae for the calculation of the LR. Similar to single-source cases, an individual with DNA profiles mixing with the victim matched with the crime trace would be identified as the suspect. However, there may be also cases in which there is no fully matched profiles found in the database. This article therefore attempts to apply the familial search to mixture cases.

Here, we present the results of an analysis of the performance of the familial search through a simulation study using Hong Kong Chinese data. Our approach is surprisingly effective in identifying the sibling or parent/child of the unknown contributor of the mixture. The use of logistic regressions to predict the power of the search is also discussed. This information is believed to be useful in guiding the decision of the police force for the initiation of crime investigations.

2. Likelihood ratio and matched loci count

In this article, we consider a two-person mixture case with DNA profile of the mixed stain found at the crime scene denoted by \( M = \{ M_l, l = 1, \ldots, L \} \), where \( M_l \) is the set of alleles at locus \( l \) present in the mixture. The observed evidences also include the DNA profiles of the victim and the set of individuals in the database \( D \), which are respectively represented by \( V = \{ V_l, l = 1, \ldots, L \} \) and \( X_0 = \{ X_{0_l}, l = 1, \ldots, L \} \). In order to identify the possible contributor to the mixture, we may first rank the database according to a specific scoring scheme and then investigate the top-listed individuals. The score assigned to each individual in the database should reflect the relatedness between the DNA profiles of the individual and the mixture, given the DNA profiles of the victim. In familial searches based on single-source stains, the scores are typically defined as a LR or a degree of allele sharing (Curran and Buckleton, 2008; Cowen and Thomson, 2008; Reid et al., 2008). When the crime trace is a two-person mixture, the hypotheses for scoring individual \( j \), \( j \in D \), using the LR approach can be formulated as follows:

- \( H_j \): the victim and the relative of individual \( j \) are contributors;
- \( H_d \): the victim and one unknown person are contributors.

Denote \( X_j = \{ X_{0_l}, l = 1, \ldots, L \} \) as the DNA profile of individual \( j \), \( P(H_j) \) as the prior probability that the unknown contributor is the relative of individual \( j \), Chung et al. (2009) derived the following general formula for computing the LR of \( H_j \) versus \( H_d \):

\[
LR_j^* = \frac{\sum_{i \in D \setminus \{j\}} P(M|V, X_j, H_j)(1 - P(H_j))}{\sum_{i \in D} P(M|V, X_i, H_i)P(H_i) + \sum_{i \in D} P(H_i)}
\]

\[
LR_j = \frac{P(M|V, X_j, H_j)}{P(M|V, H_d)}
\]

where \( H_d \) is the hypothesis that the victim and a random person contributes to the mixture and

is the ordinary LR of \( H_j \) versus \( H_d \) when only the DNA profile of individual \( j \) is available.

Using uniform priors, it can be seen from Eq. (1) that for \( j, j' \in D \), \( LR_j^* > LR_{j'}^* \) if and only if \( LR_j > LR_{j'} \). Therefore we can simply rank the database according to the values of \( LR_j \), or equivalently, \( \log(LR_j) \), for \( j \in D \). Assuming linkage equilibrium, i.e. independence of alleles across all loci, we have

\[
LR_j = \prod_{l=1}^{L} \frac{P(M_l|V_l, X_{0_l}, H_j)}{P(M_l|V_l, H_d)}.
\]

The quantities of \( P(M_l|V_l, X_{0_l}, H_j) \) and \( P(M_l|V_l, H_d) \) can be evaluated by using the \( Q \)-function in Hu and Fung (2005) and Fung and Hu (2008):

\[
P(M_l|V_l, X_{0_l}, H_j) = k_0 Q(2, U_l) + k_1 (I_{adm}(t_1)Q(1, U_l \setminus \{ t_1 \}) + I_{adm}(t_2)Q(1, U_l \setminus \{ t_2 \}))
+ k_2 I_{adm}(t_1)I_{adm}(t_2)Q(0, U_l \setminus \{ t_1, t_2 \})
\]

\[
P(M_l|V_l, H_d) = Q(2, U_l)
\]
The calculating formulae of $Q(j, B)$ for different combinations $M_i$ and $B$ at a particular autosomal locus $l$ with alleles $A_1, A_2, \ldots, A_k$ and corresponding allele frequencies $p_1, p_2, \ldots, p_k$, under the assumption of Hardy–Weinberg equilibrium. The indices $i, j, k$, and $r$ are pairwise distinct.

<table>
<thead>
<tr>
<th>$M_i$</th>
<th>$B$</th>
<th>$Q(0, B)$</th>
<th>$Q(1, B)$</th>
<th>$Q(2, B)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_i$</td>
<td>$\phi$</td>
<td>$1$</td>
<td>$p_i$</td>
<td>$p_i^2$</td>
</tr>
<tr>
<td>$A_i, A_j$</td>
<td>$\phi$</td>
<td>$1$</td>
<td>$p_i + p_j$</td>
<td>$(p_i + p_j)^2$</td>
</tr>
<tr>
<td>$A_i, A_j, A_k$</td>
<td>$\phi$</td>
<td>$1$</td>
<td>$p_i + p_j + p_k$</td>
<td>$(p_i + p_j + p_k)^2$</td>
</tr>
<tr>
<td>$A_i, A_j, A_k, A_l$</td>
<td>$\phi$</td>
<td>$1$</td>
<td>$p_i + p_j + p_k + p_l$</td>
<td>$(p_i + p_j + p_k + p_l)^2$</td>
</tr>
</tbody>
</table>

where $t_1 t_2$ is the genotype present in $X_{i1}$, $U_l = M_l \setminus V_l$ is the set of alleles present in $M_l$ but absent in $V_l$, and $I_{M_l}(t) = 1$ if $t \in M_l$ and 0 otherwise. The quantities $(k_0, 2k_1, k_2)$ are the relatedness coefficients (Fung et al., 2002) for the relationship considered in $H_l$. They respectively represent the probabilities that there are 0, 1, 2 alleles that are identical by descent between two related persons. In particular, $(k_0, 2k_1, k_2)$ take the values of $(0.25, 0.5, 0.25)$ for full siblings and $(0, 1, 0)$ for parent/child relationship. Under the assumption of Hardy–Weinberg equilibrium, for a particular autosomal locus $l$ with $K$ alleles $A_1, A_2, \ldots, A_K$ and corresponding allele frequencies $p_1, p_2, \ldots, p_K$ ($\sum_{i=1}^K p_i = 1$), the general $Q$-formula takes the form of

$$Q(k, B) = \sum_{M_l \setminus B \subseteq C \subseteq M_l} (-1)^{|M_l \setminus C|} \left( \sum_{i \in C} p_i \right)^k$$

where $B$ is any subset of $M_l$; $k$ is a non-negative integer; and $|.|$ is the cardinality of a set. It can be interpreted as the probability of $k$ alleles taken from the set $M_l$ that explain all the alleles in the set $B$. Table 1 shows the formulae of $Q(\ldots)$ for a two-person mixture.

Beside the LR approach, we also consider an alternative scoring scheme similar to the allele sharing approach. Since there can be many possible genotypes for the unknown contributor given the profiles of $(M, V)$, the number of alleles shared between each individual in the database and the contributor cannot be counted. Therefore the number of matched loci should be counted instead. We define the matched loci count (MLC) of individual $j$ as the cardinality of the set $\{l : M_l \setminus V_l \subset X_{ij} \subset M_l\}$, i.e. the number of loci at which individual $j$ has alleles matched with the $M$, after mixing with $V$. The database is then ranked according to the MLCs.

3. Simulation study

An offender database of 30 000 unrelated DNA profiles on the Identifiler™ 15 STR loci is generated according to the Hong Kong Chinese allele frequencies given in Chan et al. (2005). For each of the first 1000 profiles in the database, a related profile (full sibling or parent/child) is simulated and mixed with an independently simulated victim profile. The related profile and the mixed profile respectively represent the unobserved profile of the unknown contributor and the observed profile in the mixed crime trace. A database search is then performed using this mixture, resulting in a list of 30 000 LRs and 30 000 MLCs. The ranks of the LR and MLC of the true relative of the unknown contributor are recorded. The effectiveness of the familial search can be assessed by these ranks from the 1000 replications. For example, there are 930 replications in which the profile of the true sibling of the unknown contributor is within the top 100 profiles in the ranked list of LR, showing a 93% chance that the sibling of the unknown contributor can be successfully identified if the 100 top-listed individuals are investigated, provided that the unknown contributor is a sibling of one of the individuals in the database. Using the LR, the chance for the identification of the parent/child is more than 99%. Fig. 1 shows the probabilities of identifying the relative of the contributor in the top $k$ profiles, using LR and MLC.
As can be seen in Fig. 1, using the LR in familial search is surprisingly effective in identifying the relative of the unknown contributor even though only the DNA profiles of the mixture are observed. In roughly 90% of the cases, the sibling of the unknown contributor can be found within the top 50 profiles using LR. On the other hand, the parent/child search is more effective as investigating the top 10 profiles are already enough to identify the parent/child of the unknown contributor in 90% of the cases. Unlike the allele sharing approach for single-source case, the MLC approach provides much poorer performance than the LR approach, thereby suggesting the use of LR for familial search based on DNA mixtures. Table 2 provides additional details on the probabilities of identifying the relative of the contributor using LR.

Table 3 summarizes simulation results from various published articles on familial DNA database search using LR approach. Because of the variations in the simulation procedures, database sizes, STR systems used, and the fact that our results are obtained based on the search with DNA mixtures rather than single-source stains, a direct comparison on the effectiveness of the searches may not be very meaningful. However, the familial search of large DNA database using DNA mixture is still shown to be capable of identifying the relative of the perpetrator with false hit rate as low as in single-source cases.

It is apparent that the performance of the familial search depends much on the observed DNA profiles (M, V). If rare alleles are present in the unexplained profiles M \ V, the LR of the unknown contributor's relative will be more likely to be top listed in the database due to the extremely small random match probability P(M|V, H_R). To take into account of the observed profiles, logistic regression models were fitted to the simulated data, following the approach suggested by Cowen and Thomson (2008). For each replication of the simulation, we define a response variable θ_k that takes the value of 1 if the relative of the unknown contributor is located within the top k profiles and 0 otherwise. This response variable characterizes the outcome of a particular search that limits the investigations to the k top-listed individuals. We estimate the dependence of the outcome with respect to the observed DNA profiles by the logistic regression model.
Fig. 2. Estimated probability of identifying the (a) full sibling and (b) parent/child of the unknown contributor in the top $k$ profiles.

Table 4
Estimated parameters of the logistic regression models for the probability of identifying the relative of the contributor in the top $k$ profiles.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>$k$</th>
<th>Parameter estimate (SE)</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full sibling</td>
<td>10</td>
<td>$-5.0234(0.6557)$</td>
<td>0.3848(0.0433)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>$-4.1141(0.8317)$</td>
<td>0.3945(0.0564)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>$-3.4930(1.0170)$</td>
<td>0.3939(0.0696)</td>
<td></td>
</tr>
<tr>
<td>Parent/child</td>
<td>5</td>
<td>$-6.8409(0.8269)$</td>
<td>0.5555(0.0564)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>$-7.8720(1.1180)$</td>
<td>0.6925(0.0792)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>$-10.3790(1.6590)$</td>
<td>0.9425(0.1230)</td>
<td></td>
</tr>
</tbody>
</table>

Logit($P(\theta_k = 1)$) = $\alpha - \beta \log_{10} P(M|V, H_R)$. The coefficients of the fitted models for various values of $k$ are listed in Table 4. The p-values of all the parameter estimates are smaller than 0.001, indicating significant association between the performance of the search and the random match probability. Fig. 2 shows the estimated probabilities of identifying the sibling and parent/child of the unknown contributor within the top $k$ profiles by $-\log_{10} P(M|V, H_R)$. As can be seen, if $P(M|V, H_R)$ is less than $10^{-15}$, which is about the median of the simulated profiles, there will be a 68% chance to identify the sibling and a 99% chance to identify the parent/child within the top 10 profiles of the simulated database. This information could be used to help determine whether it is worthwhile to put effort into the investigation based on the familial search results.

4. Discussion

The use of offender databases of DNA profiles forms an important step in forensic statistics when no suspect can be identified by non-DNA evidences. Traditional familial search methods aim at identifying the relative of the perpetrator who left the single-source crime stain, when an exact match is not found in the database. In this study, we have extended the existing procedures of familial database search to include mixtures as part of the observed DNA evidence. The simulation results based on Hong Kong Chinese data demonstrated that DNA database familial search can be applied to two-person mixture cases and perform as good as in single-source cases by using the LR approach. The performance of the search depends on the database size, number of loci used, and the type of relative being searched. In particular, searching the parent/child is more effective than searching the full sibling. The other kinds of relationship are not considered because they would share less genetic similarity with the contributor and the search would be relatively less effective. The MLC approach that resembles the single-source allele sharing approach is not recommended for its high false hit rate.

The ability to find the relative of the unknown contributor of the mixture can be expressed as a predicted probability of successful identification, which is easily computed from the fitted logistic regression model using the random matched probability of the observed DNA evidences. This comprehensive measure of effectiveness can help the police force make their own decisions whether to initiate an investigation on the individuals with top-listed profiles in the database. For practical crime cases it may, however, be also necessary to determine the most appropriate scale of the criminal investigation. Therefore an important aspect of our future work must be to develop a strategy on deciding the number of individuals to be investigated after the familial search, according to the statistical criteria on the true and false hit rates required by the police force.

It is necessary, however, to point out the limitations of our familial search method presented in this article. The size of the database in our simulation study, as well as those of the simulation results compared in Table 3, are in the order of tens of thousands profiles. For large database containing millions of profiles such as the United Kingdom National DNA Database, satisfactory performance of the familial database searching is not guaranteed. Nevertheless, we believe our approach is still
applicable to many jurisdictions in which the database size is smaller than 200,000, which is the situation in most of the European countries (Hicks et al., 2009).

In addition, the presented method requires the assumption that the DNA mixture was contributed by the victim and only one perpetrator. In some criminal investigations such as group rape cases, it is possible that more than one perpetrator are involved. In such situations, the calculation of the LR could be complicated and Eq. (1) and the computational formulae presented in Table 1 are no longer applicable. Future works therefore will seek to develop more general methodology to handle the complications in the presence of multiple perpetrators.

Acknowledgements

This work was partially supported by the Hong Kong RGC General Research Fund, the Croucher Foundation and the HKU Research Output Prize Funding. The authors thank the anonymous referees for their valuable comments and suggestions that greatly improved the paper.

References