# Adapting a likelihood ratio model to enable searching DNA databases with complex STR DNA profiles

Corina C.G. Benschop[a,*], Jeroen de Jong[b], Linda van de Merwe[a] and Hinda Haned[c]
[a] Division of Biological Traces, Netherlands Forensic Institute, P.O. Box 24044, 2490AA, The Hague, The Netherlands
[b] Forensic Software Engineering Unit, Digital and Biometric Traces Division, Netherlands Forensic Institute, P.O. Box 24044, 2490AA, The Hague, The Netherlands
[c],Ahold Delhaize, Provincialeweg 11, 1506 MA, Zaandam, The Netherlands
*Corresponding author: +31 70 888 6 809, c.benschop@nfi.minvenj.nl

## INTRODUCTION

Searching a national DNA database with complex and incomplete profiles usually yields very large numbers of partial matches that can present many candidate suspects to be further investigated by the police. Current practice in most forensic laboratories consists of ordering these 'hits' based on the number of alleles matching the searched profile. However, this method does not allow differentiating candidate profiles with similar numbers of matching alleles and is subject to both high false positive and false negative errors. To address this issue and to put forward the most relevant list of candidates to the police investigators, SmartRank was developed. The SmartRank software computes a likelihood ratio (LR) with each profile in the DNA database, taking into account drop-out, drop-in and population substructure, and ranks them accordingly. SmartRank implements the LRmix model [1-3] that was adapted to enable fast and efficient searching of voluminous databases.

In this study, we describe the adaptations that were implemented to the LRmix model and compare results that were obtained by submitting several complex mixed DNA profiles to both SmartRank and LRmix Studio (lrmixstudio.org). We discuss effects of the adaptations with regard to the range of the LRs, the ranking position and software running time.

## MATERIALS AND METHODS

*DNA samples and genotype selection*

A total of 44 reference NGM DNA profiles were used of which 23 were selected from a reference set of 2,085 Dutch males [4-5]. To 10 out of the 23 selected genotypes locus SE33 was added manually, having heterozygote alleles that were observed more than once in the population frequencies as described in [5]. A further 21 genotypes were simulated using parts of the 23 genotypes, in here denoted as 'resembling donors'. Five of these 21 were generated by replacing one randomly chosen allele by a rare allele (having a frequency of 0.00024 in [5]). Twelve of the 21 genotypes were created by combining alleles from two or three of the selected 23 genotypes in such a manner that the newly simulated genotype had 100% ($n$=4), 75% ($n$=4) or 50% ($n$=4) resemblance with a two- or three-person mixture generated from the 23 existing genotypes. The remaining four genotypes were simulated so that these could potentially be from a father ($n$=2; at least one allele per locus in common) or a brother ($n$=2; sharing half of the overall profile) of one of the 23 donors.

*Simulated DNA mixtures*

Fifteen mixed DNA profiles were generated from two, three or four of the 44 reference DNA profiles, which included combinations of genotypes that show a high number or a low number of shared alleles. Eleven of the 21 resembling donors described above showed resemblance with these mixed profiles. Drop-out was applied artificially for a two- and a three-person mixed profile by randomly removing 15% or 30% of the unique alleles.

*Enhancements to the likelihood ratio model*

Q designation shut down under Hp

The first adaptation to the LR model as implemented in the SmartRank software version 0.0.11 is the shutdown of the so-called 'Q designation' under Hp. For unknown contributors under the prosecution hypothesis (Hp), and regardless of whether the evidence sample shows drop-out, the genotype combinations that are deemed possible contain alleles observed in the evidence profile. This differs from the LRmix model in which unknown contributors can have genotype combinations following the population statistics. For a collection of $m$ alleles, there are $C_{m+1}^2 = \frac{(m+1)!}{2(m-1)!}$ ways of choosing two alleles among these $m$ alleles to form a genotype. In LRmix Studio at any given locus,

$m$ is composed of all alleles observed in the population at that locus. In SmartRank $m$ is composed only of the alleles observed at this locus in the crime-scene profile. If replicates are used, $m$ contains the unique alleles.

Reducing genotype combinations for the unknown contributors under Hp results in fewer possible genotype combinations for the unknown contributors and is therefore expected to yield a higher likelihood under Hp if the candidate is a true contributor to the DNA mixture and is not subject to drop-out. If drop-out is needed to explain the profile under Hp, a lower LR is expected. For mixtures without drop-out and without allowing for drop-out in the LR calculations, the shutdown of the Q designation should not have an effect as the possible genotype combinations under Hp remains unchanged when compared to LRmix Studio.

Hd isolation

The second adaptation in SmartRank regards the defense hypothesis (Hd) calculation that is computed only once and subsequently applied in the LR calculation for every candidate profile, and independent of this candidate profile, in the DNA database. With this adaptation, the likelihood under Hd is expected to be affected if theta is non-zero. The performance gain with regard to computation time is dependent on the size of the DNA database, as Hd is calculated once instead of for each candidate in the DNA database.

*LRmix Studio and SmartRank analyses*

Mixed DNA profiles were loaded as crime scene profiles into LRmix Studio v.2.0.1 and SmartRank v.0.0.11. For each mixed DNA profile, 44 LRmix Studio LR calculations were performed by using each of the 44 reference DNA profiles as a candidate (person of interest) under Hp. Subsequently, candidates were ranked manually according to the range of the LR with the largest one on top. In SmartRank, four analyses were performed per mixed profile, namely using SmartRank 1) implementing the LRmix model without speed optimizations, 2) with the Q designation shut down for Hp, 3) with Hd isolation and 4) with both adaptations enabled. For SmartRank searches we used a DNA database comprising the 44 known reference genotypes. To compare computation times when having a larger DNA database, a DNA database was simulated to be 5,000 fold larger. This simulated database was generated based on the composition data for the Belgian and Dutch national databases and a portion of the Italian and French national databases, as well as the Dutch population statistics [5]. For all loci observed in the source databases, the probability of observing that locus in a random candidate from the database was calculated for each database separately by dividing the number of candidates containing that locus by the total number of candidates in the database. The probability of occurrence of a locus in our simulated database was determined by taking the average over the probabilities of occurrence in the source databases for that locus. When generating the simulated profiles for our database, the output of a Pseudo Random Number Generator (PRNG) was used in combination with the calculated average probability to determine the presence of any particular locus. If the locus was deemed to be present, further PRNG output in combination with the probabilities in the population statistics were used to determine the alleles. Each profile in these DNA databases was taken as a candidate under Hp and LRs for the candidate profiles were ranked by the software. Hypotheses under both the prosecution and defense hypotheses included the true number of contributors, the known drop-out rate, a theta correction of 0.01 or 0 and a drop-in probability of 0.05. For the LR computations, the Dutch allele frequencies were used [5].

## RESULTS

*Comparisons between SmartRank and LRmix Studio*

*Range of the LRs*

For comparison of the range of the LR and ranking position, a total of 2,772 LRs were calculated; 44 candidates were analyzed for a total of 15 samples, resulting in 660 LRs for LRmix Studio and 660 LRs for each of three different parameter combinations in SmartRank. Furthermore, the 44 candidates were analyzed for one of the samples using 1) SmartRank using the LRmix model without speed optimizations, 2) SmartRank with Hd isolation and no theta correction and 3) LRmix Studio without theta correction, yielding an additional 132 LRs.

First, SmartRank analyses using the LRmix model without speed optimizations were compared to LRmix Studio analyses. To that aim, a three-person mixture having 30% drop-out was taken as crime scene profile and each of the 44 reference profiles was used as candidate (person of interest) under Hp. Both software resulted in the exact same LRs for each of the 44 analyses (data

not shown), which proves that the LRmix model was correctly implemented into the SmartRank software.

As expected, shutting down the Q designation under Hp had no effect on the LRs for true donors in mixtures without drop-out (Table 1A). For mixtures with drop-out and for non-contributors, the LR tended to be lower when compared to LRmix Studio results (Fig. 1A and Table 1).

Enabling the Hd isolation mainly resulted in LRs that were within one unit on log 10 scale when compared to LRmix Studio results (Fig. 1B). With this modification larger LRs (difference more than one unit on log 10 scale) were found only for true or resembling donors, while lower LRs (difference more than one unit on log 10 scale) were obtained only for non-contributors (Fig. 1B). As expected, if theta was set to zero the LRs (*n*=44) were equal for LRmix Studio and SmartRank with Hd isolation enabled (data not shown).

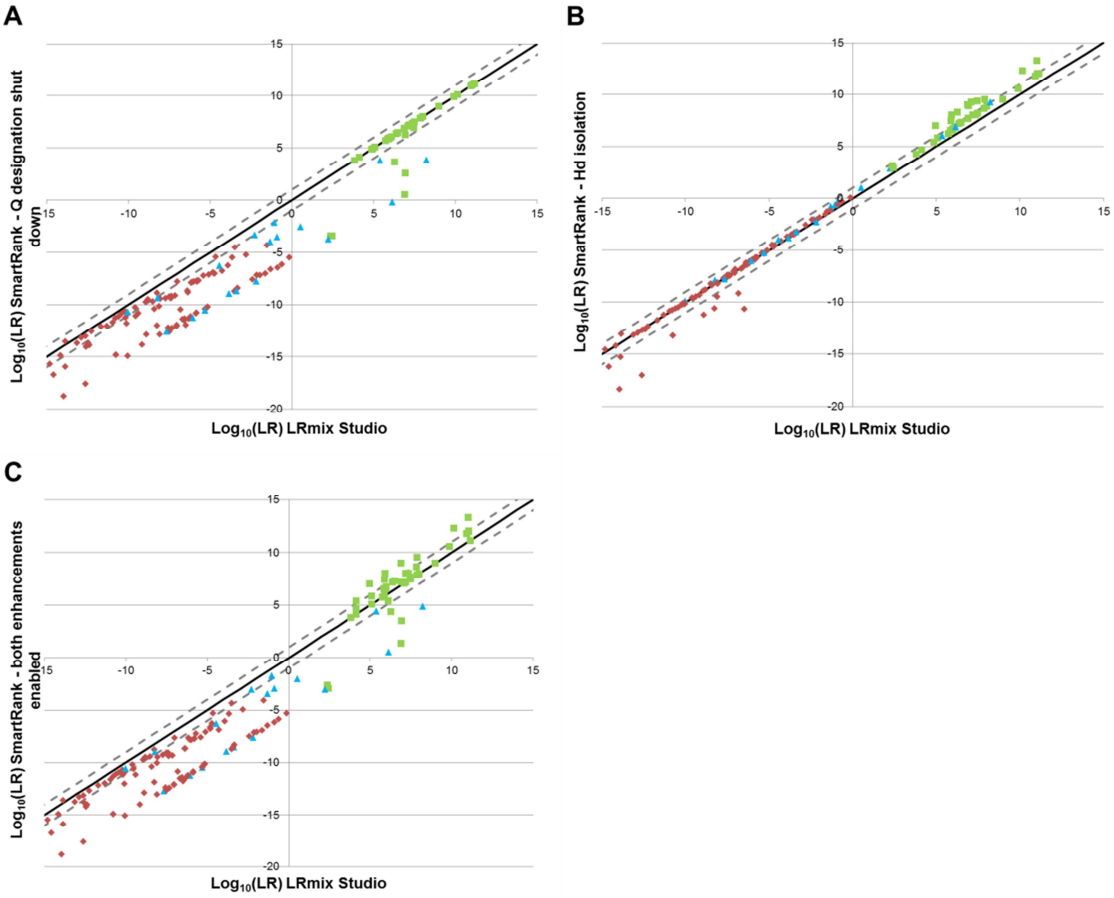


Figure 1. Comparison between LRs ($log_{10}$) obtained with LRmix Studio and (A) SmartRank with Q designation shut down for Hp, (B) SmartRank with Hd isolation and (C) SmartRank with both adaptations enabled. Green, blue and red figures correspond to results for true, resembling and non-donors, respectively. The black diagonal line represents X = Y and the dashed grey diagonal lines indicate a deviation from this line by 1 unit on the log 10 scale.

Table 1. Difference in $\log_{10}$(LR) between LRmix Studio and three different parameter combinations in SmartRank. Results for (A) true contributors, (B) resembling donors and (C) non-contributors.

| **A** | | | | Difference in $\log_{10}$(LR) compared to LRmix Studio (average and standard deviation) | | |
|---|---|---|---|---|---|---|
| Number contributors | % drop-out | # profiles (# LRs) | Q designation Hd isolation | On Off | Off On | On On |
| 2 | 0% | 4 (8) | | 0.00 ± 0.00 | 1.18 ± 0.66 | 1.00 ± 0.86 |
| 3 | 0% | 4 (12) | | 0.00 ± 0.00 | 1.12 ± 0.61 | 0.56 ± 0.72 |
| 4 | 0% | 4 (16) | | 0.00 ± 0.00 | 0.97 ± 0.54 | 0.72 ± 0.77 |
| 2 | 15% | 1 (2) | | -0.56 ± 0.11 | 1.94 ± 0.03 | 0.25 ± 0.01 |
| 2 | 30% | 1 (2) | | -3.47 ± 1.22 | 1.41 ± 0.87 | -2.63 ± 1.09 |
| 3 | 30% | 1 (3) | | -5.99 ± 0.30 | 0.69 ± 0.18 | -5.30 ± 0.28 |
| **B** | | | | | | |
| 3 | 0% | 1 (1) | | 0.32 | 0.18 | 0.18 |
| 4 | 0% | 1 (1) | | 0.00 | 0.52 | 0.52 |
| 2 | 15% | 1 (6) | | -1.02 ± 0.36 | 0.63 ± 0.66 | -0.57 ± 0.43 |
| 2 | 30% | 1 (6) | | -3.15 ± 1.01 | 0.65 ± 0.23 | -2.56 ± 0.66 |
| 3 | 30% | 1 (9) | | -5.48 ± 0.46 | 0.21 ± 0.32 | -5.25 ± 0.17 |
| **C** | | | | | | |
| 2 | 15% | 1 (36) | | -1.32 ± 1.44 | -0.67 ± 1.45 | -1.32 ± 1.44 |
| 2 | 30% | 1 (36) | | -2.22 ± 0.77 | -0.30 ± 0.86 | -2.17 ± 0.82 |
| 3 | 30% | 1 (32) | | -5.08 ± 0.22 | 0.00 ± 0.14 | -5.08 ± 0.12 |

*Ranking position*

Even though SmartRank (using the LRmix model modified to optimize speed) and LRmix Studio may yield different LRs for different parameter tunings, both software ranked all true donors within the first five positions for the samples used in this study. The non-contributors that were ranked in this top five were genotypes designed to have either 100% overlap with the mixture or differed one allele compared to one of the true contributors.

Due to the lower LRs for the three-person mixture having 30% drop-out, two out of the three donors yielded an LR below 1 when using SmartRank but not when using LRmix Studio (Fig. 1C, green figures were SmartRank yields an LR below 1). These 'false exclusions' were caused by shutting down the Q designation under Hp.

*Computation time*

For comparison of the computation time, initial analyses were performed using Java 1.7 running on a 64-bit Windows 7 computer with a dual-core Intel i5 CPU running at 2 GHz. Both LR model adaptations in itself reduced the computation time, although shutting down the Q designation hardly had an effect for mixtures without drop-out (Table 2). Hd isolation had the largest effect on the computation time. Remarkably, applying both adaptations resulted in computational times that differed only slightly from applying the Hd isolation only (Table 2).

As can be seen from Table 2, the complexity of the mixed profile has an effect on the computation time. Next to this aspect, the size of the DNA database and the type of computer that has been used influence the time required for a search. These latter two aspects were examined for a a two-person mixture with 15% drop-out and the most complex mixture in this study (a three-person mixture with 30% drop-out).

Table 3 shows the effects of using a more powerful computer, and the effects of increasing the size of the database against which the search is performed. From this table we can see that using the more powerful computer resulted in a 50% reduction of the processing time when running the simple scenario against the smaller database, and that a reduction of 82% in search time was obtained when running the more complex scenario against this database. Another observation we can make regards the effects of database size on search time when running on the faster computer. Running the simpler scenario against a 5,000 times larger database increased the search time 4,345-fold (from 2 seconds to 8,690 seconds). When the more complex scenario was run against this larger database, the search time increased 63-fold (from 215 seconds to 13,482 seconds).

Table 2. Average computation time for SmartRank with different parameter combinations, a DNA database comprising 44 reference profiles and using Java 1.7 running on a 64-bit Windows 7 computer with a dual-core Intel i5 CPU running at 2 GHz.

| Number of contributors | % drop-out | # profiles (# LRs) | SmartRank running time (average) | | | |
|---|---|---|---|---|---|---|
| | | | Q designation<br>Hd isolation | Off<br>Off | On<br>Off | Off<br>On | On<br>On |
| 2 | 0% | 4 (8) | | 0:00:04 | 0:00:04 | 0:00:01 | 0:00:01 |
| 3 | 0% | 4 (13)[a] | | 0:03:14 | 0:03:08 | 0:00:05 | 0:00:05 |
| 4 | 0% | 4 (17)[b] | | 03:12:53 | 3:26:25 | 0:05:03 | 0:05:00 |
| 2 | 15% | 1 (44) | | 0:03:47 | 0:01:28 | 0:00:05 | 0:00:04 |
| 2 | 30% | 1 (44) | | 0:01:34 | 0:01:26 | 0:00:05 | 0:00:04 |
| 3 | 30% | 1 (44) | | 15:57:27 | 14:30:26 | 0:20:47 | 0:19:55 |

[a] All 12 true contributors and one resembling donor, that has 100% of its alleles overlapping with the mixture, yielded a LR.
[b] All 16 true contributors and one resembling donor, that has 100% of its alleles overlapping with the mixture, yielded a LR.

Table 3. Computation time for SmartRank v.0.0.11 with both adaptations enabled, run on two different systems and using a DNA database (DDB) of size 44 or 220,000.

| Number of contributors | % drop-out | # profiles (# LRs) | DDB size 44 | | DDB size 220,000 |
|---|---|---|---|---|---|
| | | | Intel i5 [a] | Intel Xeon [b] | Intel Xeon |
| 2 | 15% | 1 (44) | 0:00:04 | 00:00:02 | 02:24:50 |
| 3 | 30% | 1 (44) | 0:19:55 | 00:03:35 | 03:44:42 |

## DISCUSSION AND CONCLUSIONS

This study shows that the LRmix model was correctly implemented in the SmartRank software as LRs were identical when using the same model parameters. The SmartRank software implements two adaptations to reduce computation time enabling efficient computation of LRs for large DNA databases. These adaptations reduced the computation time by 3 seconds for simple two-person mixtures without drop-out up to over 15 hours for a complex three-person mixture with drop-out when using a small DNA database of 44 profiles. As expected, using a 5,000 fold larger DNA database negatively influences the computation time, while the use of a more powerful computer positively affects the time that is required for a SmartRank analysis. Applying the speed optimizations comes at cost of deviations in the range of the LRs when compared to LRmix Studio results (Fig. 1). Despite these deviations, all true contributors in this study were ranked within the top five list of candidates. Such a list could serve a useful investigative lead in criminal cases lacking a suspect, and fulfills the goal of SmartRank. However, for the most complex sample in this study (three persons with 30% drop-out), the Q designation shut down under Hp resulted in an LR below 1 for two of the three true contributors, which was not the case when using LRmix Studio. Future developments include the implementation of frequency coalescing to further reduce computation time, which may lead to redundancy of the Q designation shut down under Hp and obtaining more true positives. At current, we are conducting a large validation study, yielding guidelines for best practice and further insight in the applicable domain of the software. These results will be described elsewhere.

## ACKNOWLEDGEMENTS

This study received funding support from the European Network of Forensic Science Institutes (ENFSI) Monopoly programme “SmartRank: A likelihood ratio software for searching national DNA databases with complex DNA profiles” (MP2013-T4).

## REFERENCES

[1] H. Haned, P. Gill, Analysis of complex DNA mixtures using the Forensim package, Forensic Sci. Int. Genet. Suppl. (2011) e79–e80.
[2] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (2011) 265–268.
[3] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, Forensic Sci. Int. Genet. 6 (2012) 762–774.

[4] A.A. Westen, H. Haned, L.J. Grol, J. Harteveld, K.J. van der Gaag, P. de Knijff, T. Sijen, Combining results of forensic STR kits: HDplex validation including allelic association and linkage testing with NGM and identifiler loci, Int. J. Legal Med. 126 (2012) 781–789.
[5] A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, K.J. van der Gaag, N.E.C. Weiler, J. Warnaar, M. Kayser, T. Sijen, P. de Knijff, Comparing six commercial autosomal STR kits in a large Dutch population sample, Forensic Sci. Int. Genet. 10 (2014) 55–63.